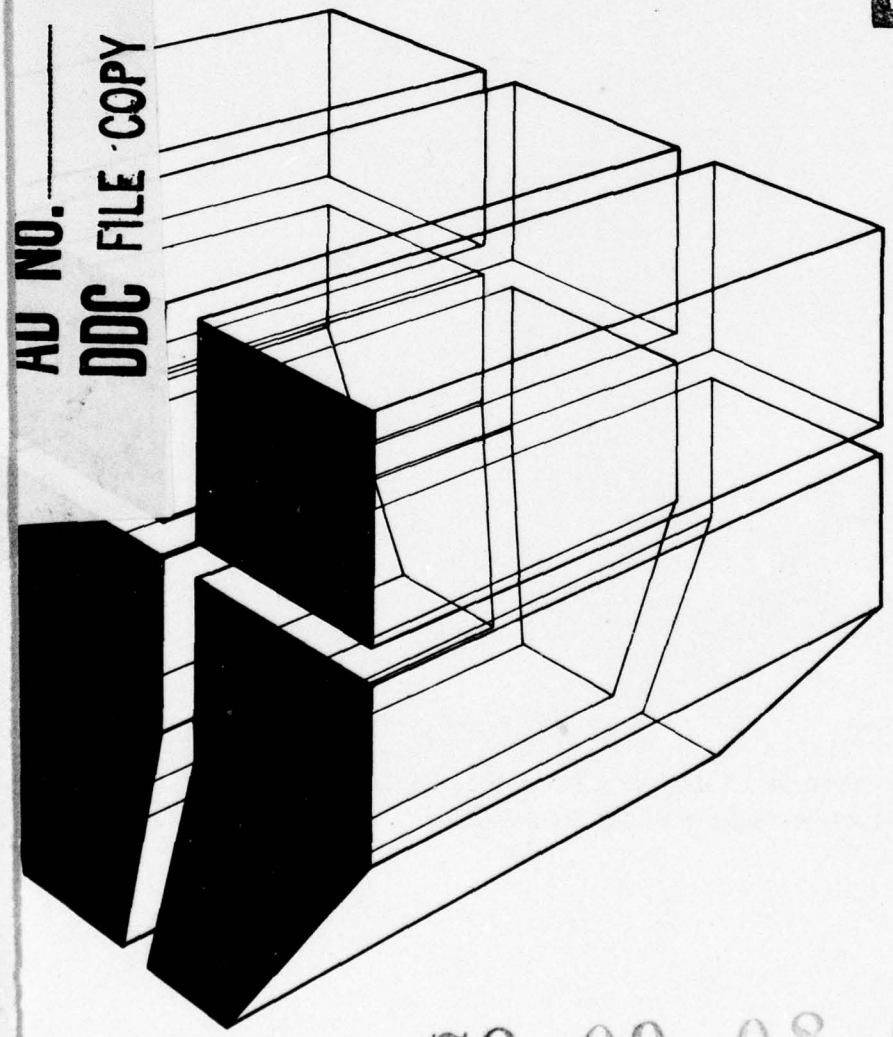
construction
engineering
research
laboratory

SPECIAL REPORT E-132
August 1978

USE OF "IDEAL" RATINGS AS A STANDARD FOR EVALUATING FACILITIES

LEVEL II



by
Wayne D. Veneklasen
Roger L. Brauer
Bruce Sevy

CERL

The contents of this report are not to be used for advertising, publication, or promotional purposes. Citation of trade names does not constitute an *official indorsement or approval of the use of such commercial products.* The findings of this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

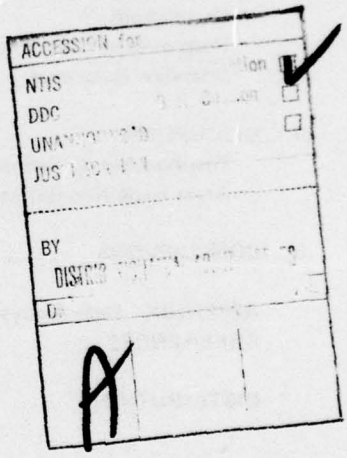
DESTROY THIS REPORT WHEN IT IS NO LONGER NEEDED
DO NOT RETURN IT TO THE ORIGINATOR

FOREWORD

This research was conducted for the Directorate of Military Construction, Office of the Chief of Engineers (OCE), under Project 4A161102AT23, "Structure Systems"; Task 01, "Facility System Performance"; Work Unit 002, "Semantic Scales as Standards for Evaluating Facilities." The applicable QCR is 1.01.012.

The work was performed by the Energy and Habitability Division (EP), U.S. Army Construction Engineering Research Laboratory (CERL), Champaign, IL. The Principal Investigator was Dr. Roger L. Brauer. Dr. Wayne D. Veneklasen and Mr. Bruce Sevy were Associate Investigators. Mr. R. G. Donaghy is Chief of EP.

COL J. E. Hays is Commander and Director of CERL and Dr. L. R. Shaffer is Technical Director.



CONTENTS

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: CERL-SR-E-132	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): USE OF 'IDEAL' RATINGS AS A STANDARD FOR EVALUATING FACILITIES.	5. TYPE OF REPORT & PERIOD COVERED: SPECIAL Report
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Wayne D. Veneklasen, Roger L. Brauer, Bruce Sevy	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: CONSTRUCTION ENGINEERING RESEARCH LABORATORY, P.O. Box 4005, Champaign, IL 61820	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 4A161102AT23-01-002 07
11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE: August 1978
	13. NUMBER OF PAGES: 37
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited. 39 p.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES
Copies are obtainable from National Technical Information Service
Springfield, VA 22151

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report presents the results of a study conducted to determine the reliability and validity of using 100-mm bipolar semantic scales to establish an ideal which can be used as a standard for subjectively evaluating facilities. Such an ideal standard would permit all facilities to be evaluated by looking at the difference between the profiles of ratings of existing and ideal facilities.

Data were obtained through questionnaires administered to Air Force enlisted personnel, Army officers, and civilian Army office workers. Items rated using the 100-mm scales

→ over

Block 20 continued.

were existing and ideal dining facilities, existing and ideal Bachelor Officers' Quarters (BOQs), an ideal BOQ entrance, and an ideal wristwatch. Additional data were obtained through interviews.

Results indicate that the 100-mm bipolar rating scales could be a viable evaluation tool with one qualification: to provide any meaningful evaluation, the ideal scales must be paired with some dependent measure, such as existing scales. Without such a basis for comparison, there is very little differentiation between ideal ratings of various objects. The data indicate, in fact, that the 100-mm technique itself may influence a person's response to a greater extent than the type of object being rated does.

TABLES

FIGURES

USE OF "IDEAL" RATINGS AS A STANDARD FOR EVALUATING FACILITIES

1 INTRODUCTION

Problem Statement

Built facilities often need to be evaluated; i.e., their worth or quality must be judged or determined. The term "facility evaluation" has several shades of meaning. Facilities may be evaluated for many different reasons, by many different people, in many different ways (see the *Types of Evaluation* discussion in the Background section). Complete evaluation of a facility often requires obtaining subjective information from occupants. One reason occupants or users of facilities have not been involved in facility evaluation more frequently is that the use of subjective information requires a standard for comparing one group of occupants to another or one facility to another. This study focuses on one approach to this problem: use of 100-mm bipolar adjective scales to define an "ideal" against which facilities can be evaluated.

Objective

The objective of this study was to determine the reliability and validity of using 100-mm bipolar semantic scales to establish an ideal which can be used as a standard for evaluating facilities. Establishment of an ideal standard would permit all facilities–new or old, existing or proposed–to be evaluated by looking at the difference between the profiles of ratings of existing and ideal facilities. This report describes the study method and results and outlines cautions to be exercised in using this approach.

Approach

The study was conducted in the following steps:

1. Data collected for previous U.S. Army Construction Engineering Research Laboratory (CERL) studies of Air Force dining facilities were retrieved and reanalyzed.

2. New data on another facility type were collected in questionnaire form, and other, nonfacility objects were rated using the scales used to evaluate facilities.

3. Interviews were conducted to determine the referent used by the respondents when they marked the bipolar scales. For example, it was necessary to know whether respondents were referring to such things as natural or artificial light when they evaluated lighting. This information was collected for all 10 of the 100-mm bipolar scales used in the study.

4. The previously collected dining facility data were subjected to a computer screening prior to statistical analysis. This screening eliminated all responses not meeting the criteria for inclusion. The screening included checks for within-score consistency (to eliminate random responses) and missing and illegal responses (responses not contained within the 100-mm range).

5. Test-retest reliability for the ideal scales was measured in a laboratory setting.

6. Statistical analyses were performed on the data to estimate validity and reliability. The scales were then compared across usages.

Background

Confusion regarding what facility evaluation is or means frequently arises. To provide greater insight into facility evaluation, this section briefly discusses the types of facility evaluation and who performs the evaluation.

Types of Evaluation

Evaluation of facilities can be classified in many ways. One way is on the basis of *subject matter*. For example, facilities can be evaluated on economic grounds; economic considerations may include initial cost or the cost of operation and maintenance. Facilities can also be judged based on their quality, usually in terms of physical characteristics. These physical characteristics can also be subdivided in many ways. For example, they could be classified as environmental conditions, functional aspects, subsystems of the facility (e.g., lighting), and subsystem components (e.g., a switch or a lighting fixture).

Another way of classifying evaluation of facilities is on the basis by which *judgments are formed*–objectively or subjectively. In forming objective judgments, existing conditions or features are compared to some standards or limits which are expressed concretely and are universally understood. In forming subjective judgments, existing features or conditions are compared to internal standards unique to each judge.

A third method by which facility evaluation can be classified is on the basis of *what is being evaluated*–

existing facilities or a potential facility described in drawings or other forms. In Military Construction, Army (MCA) procedures, various regulations describe several types of evaluation for existing facilities (Table 1).

Table 1
Types of Evaluations of Existing Facilities

Source of Procedure	Type of Evaluation
AR 210-20	Evaluating existing facilities for adequacy to support the mission
ER 415-3-11, par. 4	Post-completion inspection (after facility has been in use 6 months)
ER 415-3-11, par. 6	Criteria feedback evaluation (after facility has been in use 3 years)
ER 415-345-38, par. 3	Inspection at transfer of completed facility

Similarly, evaluation of potential facilities is governed by AR 415-20 (Design Approval). Two procedures resulting from this are design analysis (required by ER-1110-345-700) and design verification (required by ER-1110-345-100, par. 17).

A fourth way of classifying facility evaluations is based on the *purpose for conducting the evaluation*. If (1) distinctions are made between requirements, criteria, specifications, and design solutions, and (2) requirements are the basis for criteria, criteria are the basis for specifications, and both criteria and specifications are the basis for design solutions, then each of these items requires a different type of evaluation. The confusion arises when it is recognized that the same facility may be used for evaluating each of these items. Since, for example, evaluation of the design solution may require a different approach than evaluation of the criteria used in achieving the solution, the purpose of evaluating the facility must be known. (Criteria evaluation was discussed in CERL Special Report D-78.[1])

In summary, categorizing the evaluation of a facility may involve all four systems of classification. For example, an existing facility may be used to evaluate the effectiveness of some criteria using objective measures of physical characteristics. In another case, there may be a need to evaluate a physical characteristic of a design solution in an existing facility using the subjective views of occupants in order to understand its significance for occupant satisfaction.

Involvement of Occupants in Building Evaluation

Complete evaluation of a facility requires many different tools; in some cases, the tools may involve the occupants. Occupants can provide valuable input in building evaluation in several ways. First, they can report on the physical aspects of a facility. They can and frequently do report whether a subsystem or one of its components is inoperative. They can also report on functional aspects of a facility, describing which things work well for them and which do not. In addition, they can provide views about the many qualitative features or characteristics of a facility which can only be assessed through subjective evaluation.

For example, to make decisions about such subjective things as the impact of a facility on the morale and satisfaction of occupants, information about the opinion of a group of occupants must be available. Such evaluations can be useful and are technically feasible. A family housing study showed that as much as 60 percent of the variance in the overall satisfaction of occupants with their housing could be accounted for by their ratings of *specific* interior features.[2]

The "Ideal" as a Subjective Standard

As previously stated, use of subjective occupant evaluation creates a problem—the need for a standard for comparing groups or facilities. While economic and physical factors have some sort of standard against which evaluations can be made, there are usually no standards of evaluation for the subjective input. Most user input obtained involves attitudinal information based on individualized value systems; each individual's own set of standards determines his/her behavioral responses and attitudes about the environment.

Attitudinal, or judgmental, evaluation represents a large percentage of the reported research of man-environment relations and building evaluation. Products of such research usually take the form of factor-analytic *descriptions of the data* or analyses of variance which account for *variance in the data*. In either case, the

[1] R. L. Brauer and D. L. Dressel, *Concepts for the Generation, Communication, and Evaluation of Habitability Criteria*, Special Report D-78/ADA041187 (U.S. Army Construction Engineering Research Laboratory [CERL], 1977).

[2] D. L. Dressel et al., *Army Family Housing: Preferences and Attitudes about Housing Interiors, Vol III: Predictors of Satisfaction with Housing Interiors*, CERL Technical Report D-48/ADA011187 (CERL, April 1975).

results describe the data rather than the building being evaluated. There have been no major criteria-development studies that have generated sets of criteria or standards against which attitudes can be compared.

A previous CERL research effort examined the possibility of establishing a standard for attitudinal information. In that research, Air Force enlisted personnel at Travis AFB were asked to rate their existing dining halls and an ideal dining hall measuring up to their individual standard of "ideal" using a set of eleven 100-mm bipolar adjective scales. The results were organized as profiles of the ratings of the dining halls. A comparison of the profiles indicated that there were statistically significant differences between the ratings of the existing dining halls, but that ratings of an ideal dining hall were essentially the same.[3]

To further examine the concept of an ideal, CERL investigated the generalizability of the ideal to Air Force dining halls at posts in different geographic locations and having varying missions. Results of this second study, which was conducted at Minot AFB, ND, Homestead AFB, FL, and Travis AFB, CA, indicated that the ratings of an ideal dining hall were the same at all three locations, while ratings of the existing facilities were again statistically different.[4]

Use of the ideal as a standard of evaluation was investigated in CERL's first study at Travis AFB. Pre- and post-renovation data were collected on the existing dining halls and an ideal dining hall. Over time, the ideal ratings did not change for either the experimental group (renovation) or the control group (no renovation), indicating that the concept for the ideal did not change either over time or from group to group. The ratings of the existing dining halls after renovation did change significantly, moving closer to the ideal ratings.

Conclusions from both CERL studies generally indicated that the use of an ideal may have some generalizability as a standard for the evaluation of dining halls. No data were collected on other facility types, but the use of the ideal certainly proved worthy of further investigation to demonstrate the reliability and validity of the concept. If that demonstration can be completed, the result would be at least one standard that can be used to evaluate buildings using attitudinal data. A valid and reliable bipolar medium for obtaining the ideal standard would be a fast, efficient, and inexpensive way of obtaining quantifiable user input to the building design and delivery process. Rating profiles could also be used to evaluate existing facilities before and following renovation by using the ideal as the methodological control.

[3]W. Gibbs, *Comparison of Consumer Satisfaction Before and After Dining Facility Renovations at Travis AFB, CA*, Technical Report D-28/AD784056 (CERL, 1974).

[4]W. Gibbs, *Comparative Study of Consumer Attitudes at Three Air Force Dining Facilities*, Interim Report D-40/ADA-000711 (CERL, 1974).

2 METHOD

Administration of Scales

The 10 bipolar descriptive scales shown in Table 2 were administered by questionnaire to 868 enlisted Air Force personnel, 287 Army officers, and 49 civilian Federal office workers at CERL. Table 3 presents a breakdown of the respondents by location. The scales used were identical in all situations, except that the scales administered to the Air Force personnel (as part of a larger questionnaire in a previous CERL study) included an additional usual/unusual pair.

Table 2
Scales Used

Brightly Lighted	Dimly Lighted
Noisy	Quiet
Crowded	Uncrowded
Ugly	Beautiful
Drab	Colorful
Unpleasant	Pleasant
Cluttered	Uncluttered
Uninviting	Inviting
Run Down	Well Kept
Poorly	Well Organized

Uses of Scales

The 10 bipolar scales were used in a variety of ways. First, Air Force enlisted personnel rated their existing and ideal dining facilities using the scales. These data from an earlier CERL research effort were simply retrieved and re-examined. The 287 Army officers were asked to use the scales to rate their existing and ideal Bachelor Officers' Quarters (BOQs).

To determine whether the bipolar scales were measuring what they were intended to measure (i.e., to test their validity), further data were collected from a subsample of the officers surveyed at Fort Bliss. This subsample of 53 officers rated an ideal BOQ entrance (in

Table 3
Breakdown of Sample by Installation

Service	Location	Number of Respondents
Air Force	Travis AFB, CA	614
	Minot AFB, ND	145
	Homestead AFB, FL	109
Army	Fort Lewis, WA	19
	Fort Bliss, TX	112
	Fort Meade, MD	31
	Fort Leonard Wood, MO	64
	Fort Benjamin Harrison, IN	61

addition to their existing and ideal BOQs) to permit observation of the scales' capability to discriminate by facility type.

To test the scales' capability to discriminate between ratings of facility types and a "nonsense" item completely unrelated to buildings, 49 Federal civilian employees were asked to rate an ideal wristwatch using the scales.

The last use of the scales involved two administrations to 24 CERL personnel to determine the scales' test-retest reliability. The respondents were asked to rate their existing and ideal offices; the scales were administered to the same sample again, 5 weeks after the first administration.

Data From Scale Administrations

On all scales, the data represented the measured distance, in millimeters, from the negative descriptor (which was zero) to the respondent's actual evaluative mark on the 100-mm line. The data can be directly interpreted into percentage figures based on 100 equal units of measurements in millimeters.

All answers were measured with a 100-mm ruler and recorded on computer layout sheets. The data were then keypunched and the cards were used for the various statistical analyses.

Data Screening

To determine how "clean" (free of invalid responses) the data were, the responses from the Air Force personnel were subjected to a computerized screen* which analyzed each set of responses. Only the Air Force data were used, because the large sample size most closely fit the requirements of the computer program. The data were first screened to determine the number of illegal response patterns (those outside the 100-mm range) and missing responses. The cleaned data were then analyzed for within-score consistency (WSC) to eliminate random responses. The screening procedure is described more fully in the appendix.

Data Analysis

Initial descriptive statistics and histograms were computer-generated to permit empirical examination of the data. The primary focus of the statistical analyses was placed on the scales themselves and comparisons between them by usage. First the existing and ideal scales used on BOQs and dining halls were analyzed. Once this analysis stage was finished, the scales were compared across usages. For the actual statistical analyses, the ONEWAY analysis of variance program (ANOVA) from the Statistical Package for the Social Sciences (SPSS) was used along with Duncan's Multiple Range Test and Scheffé's test to determine mean differences.

Additional Data Collected

To obtain a better understanding of how the bipolar pairs were being interpreted, a second subsample of the officers surveyed at Fort Bliss was selected. These 41 officers were interviewed regarding the referents used when they answered the questionnaire. A separate question was asked regarding each of the 10 bipolar scales used in assessing ideal and existing environments, and all responses were content-analyzed.

Additional data were also obtained by asking 29 Army questionnaire respondents to answer an open-ended question appearing at the end of the questionnaire booklet; the respondents were asked to describe in their own words what an ideal BOQ would be like.

*The computer program used was developed by an independent research organization, the Institute for Behavioral Research in Creativity, Salt Lake City, UT.

3 RESULTS

This chapter summarizes the results of the data analyses and presents results of the open-ended questions, interviews, and reliability checks. Details of the data analyses, which were performed by a private contractor, are presented in the appendix.

Ratings of Existing BOQs

A ONEWAY ANOVA of the ratings of existing BOQs showed that there were significant differences between Forts Bliss, Lewis, Meade, and Leonard Wood on all of the 10 bipolar scales except three—noise, clutter, and organization. However, as Figure 1 shows, even in those cases where there was a significant difference between scales, the spread of scores on any given scale was not very large. The maximum difference between any two rankings was 17 mm. Even if the 100-mm line were to be broken into seven discrete categories, the greatest difference between any two rankings would be no more than one category.

Further analysis using the Scheffé Multiple Range Test indicated which forts were significantly different from others on any given scale. These results are illustrated in the second column of Figure 1. Note that six possible paired-post comparisons can be made for all four installations. In no case did more than two of the six comparisons result in significant differences.

The relative scarcity of significant differences among the comparisons of existing conditions suggests that the four forts are basically similar. Even in those cases where there was a significant difference between two forts, the absolute difference involved a maximum of 20 percent of the 100-mm line. Figure 1 shows the mean scores for each fort on each scale. The scores are grouped fairly tightly and tend to stay slightly below the 50-mm point, with the lowest ranking on any scale being 27 mm and the highest ranking 60 mm. Thus, BOQs are seen by their occupants as being generally, though not extremely, on the low or negative side of average.

Ratings of Ideal BOQs

The same bipolar scales were used to assess the officers' concept of an ideal BOQ. A ONEWAY ANOVA showed that there was no difference between forts in the officers' concept of an ideal BOQ. For each scale shown in Figure 2, the mean scores tended to stay close to 80 mm, with the highest ranking on any scale being 91 mm and the lowest being 65 mm. The fact that no significant differences were found between the four forts on any scale may mean one of two things: (1) that officers' concept of an ideal BOQ is quite similar no matter where they are stationed, or (2) that the scores are artifacts produced by a response set that is inherent in the 100-mm technique.

It is interesting to note that on the average the scores on the brightly lit/dimly lit and cluttered/uncluttered scales tend to be lower than the scores on the other eight scales. As presented to the subjects, the positive side of each scale was to the right of the 100-mm line for all scales except the two previously mentioned. This suggests that there may have been some contamination of the scores caused by a response set such that the marking of the right scales influenced the marking of the test. When the subjects came to a scale where the polarity was reversed, they may have had to stop and reconsider the 100-mm line independently of their previous responses. However, as they returned to adjacent scales of similar polarity, the visual stimulus of the previous rating may have exerted a stabilizing influence that tended to damp out any radical variation among responses.

Comparison Across Facility Types

Since there was no difference between the four forts on any of the ideal scales, the data were collapsed across forts to form a grand mean for each of the 10 scales. The same 10 scales were then used to assess 53 officers' concept of an ideal BOQ entrance and 49 CERL office workers' concept of an ideal wristwatch. The scores for the latter two concepts were compared to the grand means for the ideal BOQ scales. A ONEWAY ANOVA of these three sets of scores showed that there were significant differences on six of the 10 scales. A Scheffé Multiple Range Test was then used to specify where the exact differences lay. Figure 3 shows the differences between groups for each of the 10 scales. These data further complicate the interpretation of the 100-mm technique. The fact that four of the scales show no significant differences between groups suggests that the ideal will be constant no matter what is being rated. Whether this is because of an insensitivity inherent in the 100-mm technique or because the bipolar adjectives used were irrelevant to the object being rated is impossible to determine from the data that have been collected. On the other hand, six of the scales did discriminate between groups, although not every group was significantly different from every other group for each scale. This suggests that the 100-mm technique is somewhat viable as a tool capable of assessing the concept of the ideal.

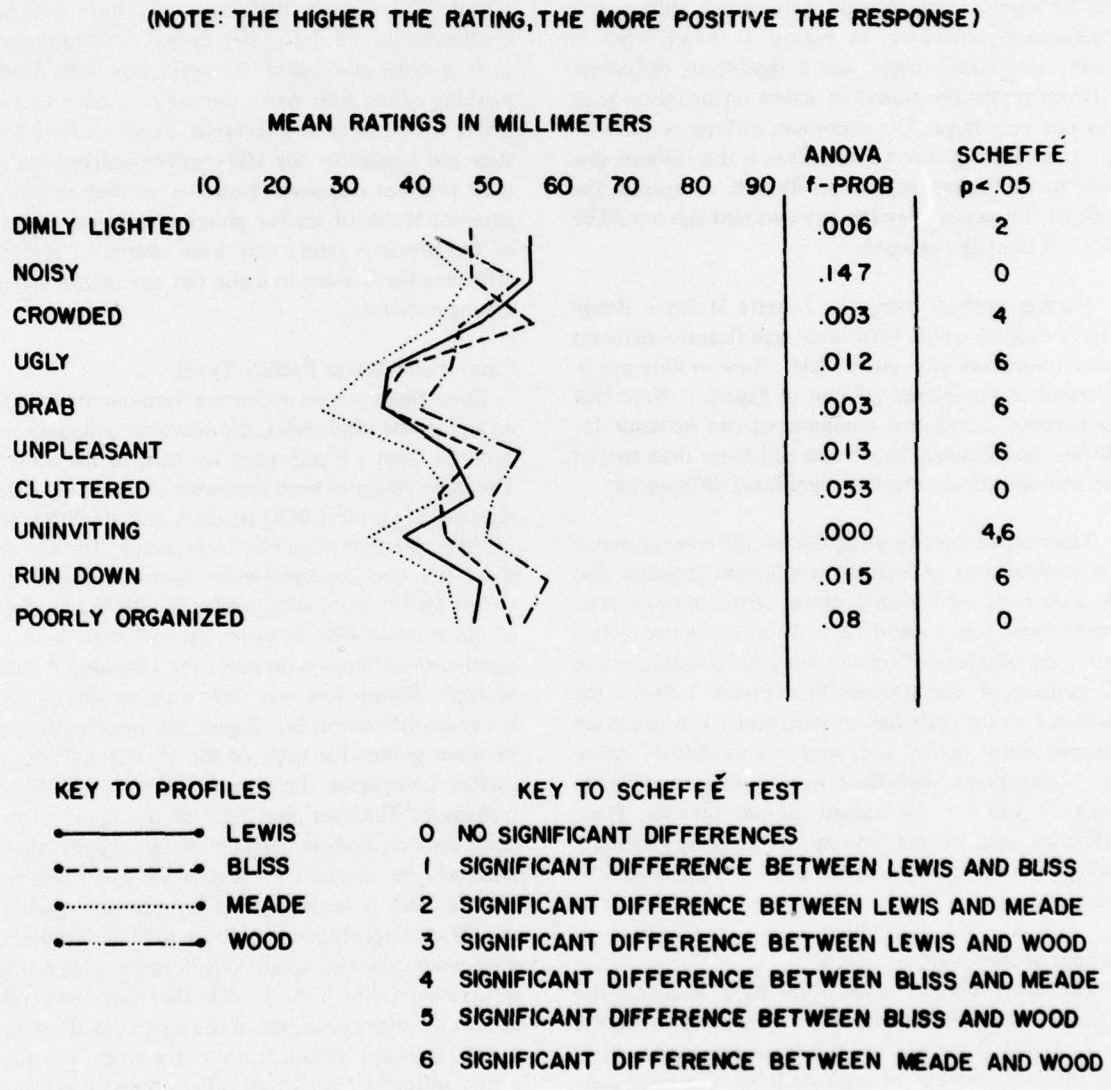
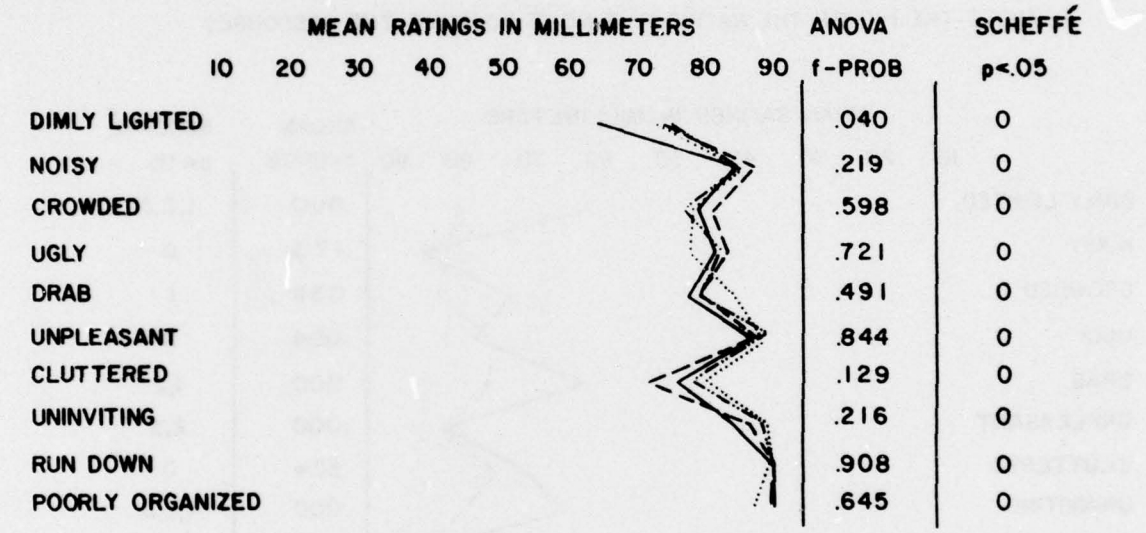


Figure 1. Profiles of existing BOQs.

(NOTE: THE HIGHER THE RATING, THE MORE POSITIVE THE RESPONSE)



KEY TO PROFILES

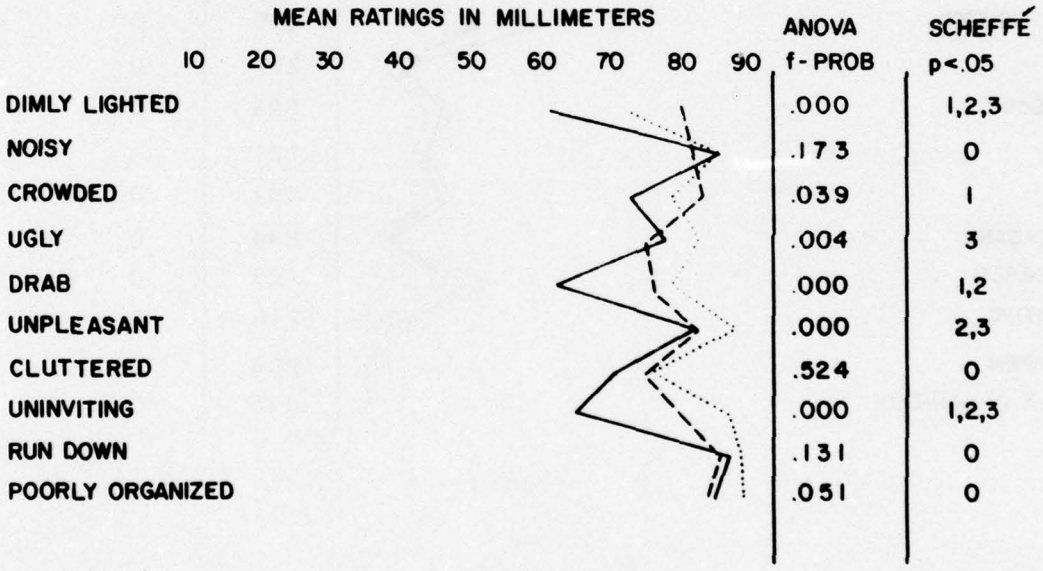
- LEWIS
- BLISS
- MEADE
- WOOD

KEY TO SCHEFFÉ TEST

0 NO SIGNIFICANT DIFFERENCES
1 SIGNIFICANT DIFFERENCE BETWEEN LEWIS AND BLISS
2 SIGNIFICANT DIFFERENCE BETWEEN LEWIS AND MEADE
3 SIGNIFICANT DIFFERENCE BETWEEN LEWIS AND WOOD
4 SIGNIFICANT DIFFERENCE BETWEEN BLISS AND MEADE
5 SIGNIFICANT DIFFERENCE BETWEEN BLISS AND WOOD
6 SIGNIFICANT DIFFERENCE BETWEEN MEADE AND WOOD

Figure 2. Profiles of ideal BOQs.

(NOTE: THE HIGHER THE RATING, THE MORE POSITIVE THE RESPONSE)



KEY TO PROFILES

- CERL
- ENTRANCE
- BOQ'S

KEY TO SCHEFFÉ TEST

0 NO SIGNIFICANT DIFFERENCES
1 SIGNIFICANT DIFFERENCE BETWEEN CERL AND ENTRANCE
2 SIGNIFICANT DIFFERENCE BETWEEN CERL AND BOQ'S
3 SIGNIFICANT DIFFERENCE BETWEEN ENTRANCE AND BOQ'S

Figure 3. Profiles of all ideal ratings.

Correlations

All 10 scales of the three groups pertaining to existing BOQs, ideal BOQs, and ideal entrances were correlated within and between each group. Typically, the only significant correlations (> .73) were found between scales within a given group. There were no significant correlations on any scales between existing and ideal BOQs, although there were a number of significant correlations between ideal BOQs and ideal entrances. As would be intuitively apparent, within-group scales such as beautiful, colorful, pleasant, inviting, etc., tended to correlate highly. This largely reflects the fact that all responses on a given scale tended to be fairly closely grouped, as previously mentioned. The fact that there were no significant correlations between scales in the existing and ideal groups suggests that the existing conditions of a respondent's quarters do not influence his or her perception of what ideal quarters would be like. The relatively large number of correlations between the ideal BOQs and ideal entrances groups is to be expected given the previous observation that ideal BOQs, ideal entrances, and ideal wristwatches tended to be rated quite similarly.

Some further observations can be made from the correlational data. Most significant of these is the fact that relatively more significant correlations occur with beautiful, colorful, and pleasant than with any of the other scales. Typically, these three correlate most strongly with the last six scales on the form, with the exception of the cluttered/uncluttered scale. This correlation seems reasonable, given the connotative similarity of the individual scales. One would expect well kept to correlate with pleasant, but there is no obvious intuitive reason to expect brightly lit to correlate with, say, pleasant or colorful. In this sense, then, the correlational data support an intuitive judgment of the similarity of the various scales. This support in turn lends credence to the use of semantic differential scales in general and suggests that the trouble with the 100-mm technique lies more with the instrument itself than the concept of the ideal.

Responses to Open-Ended Question

Responses to the open-ended question showed that six categories tended to dominate the responses: space, furnishings, storage, decor, environment, and privacy. The twenty-nine respondents generated 137 comments describing their concept of an ideal BOQ. Table 4 shows the number of comments and percentage of responses per category.

Table 4
Responses to Open-Ended Question

Category	Number of Comments	Percentage
Space	28	20.3
Furnishings	41	30
Storage	6	4.3
Decor	24	17.4
Environment	21	15.2
Privacy	12	8.7
Other	5	3.6
Total	137	99.5

In the space category, the major descriptors used were amount and arrangement of space, followed by comments on specific areas such as kitchen, bath, and living. Furnishings were described by style, comfort, and color. Quantity and location defined storage, while personalization, color, and materials defined decor. Environment was described by lighting, temperature, pleasantness, inviting atmosphere, upkeep, and noise. The last major category, privacy, was described by living privacy (intrusion by numbers of people, noise intrusion) and by private entry to the BOQ room.

It is interesting to note that the 10 scales that make up the 100-mm scale are generally represented by at least one of the general categories or subcategories. Unfortunately, it is hard to determine how much of this overlap was caused by the respondents' previous exposure to at least two 100-mm scales which suggested at least the broad areas of interest to the researchers. However, the officers' responses were often much more specific than the bipolar adjectives used on the 100-mm scale, suggesting that even if the responses were suggested by the 100-mm scales, the respondents felt that there was a need for additional specificity in the wording of the items. For example, in the open-ended question, color and style were mentioned under both decor and furnishings, whereas in the 100-mm scales colorful/drab and ugly/beautiful were related to BOQs in general rather than any one specific area.

Interview Responses

Results of the interviews with the subsample of 41 officers (Table 5) showed that none of the bipolar pairs were uniformly interpreted by all respondents. Only one of the 10 scales (cluttered/uncluttered) was interpreted in less than three ways, with the majority of the pairs being interpreted in at least five ways. While the interviews illustrated that there was a wide

Table 5
Interview Response Tabulation

Scale	Interpretations	Number
Brightly/Dimly Lighted	Natural Light	7
	Artificial Lighting	13
	Adequacy	5
	Intensity	1
	Variability	8
	Existing Combination of Both Natural and Artificial	1
Noisy/Quiet	Conversation	4
	Stereo/TV	5
	Neighbors	18
	Interior (fans, toilets, etc.)	7
	Exterior (parking lot, lawn mowers)	14
Crowded/Uncrowded	Furniture	23
	People	9
	Storage	2
	Floor Space	5
Ugly/Beautiful	Color	13
	Furniture	10
	Decor	5
	Cleanliness	2
	Style	2
	Walls	5
	Personal Belongings	3
Drab/Colorful	Color	21
	Walls	22
	Rugs/Carpet	5
	Furniture	6
Unpleasant/Pleasant	Color	6
	Furnishings	8
	Atmosphere	7
	Temperature	3
	Comfort	4
	Personal Effects	4
Cluttered/Uncluttered	Personal Effects	9
	Furnishings	19
	Other	14
Uninviting/Inviting	Pride in BOQ	7
	Cleanliness	1
	Building Layout	3
	First Impression	6
	Atmosphere	6
	Arrangement of Furnishings	4
	Other	4
Run Down/Well Kept	Walls	11
	Maintenance	14
	Maid Service	6
	Equipment	7
	Decor	5
	Interior and Exterior	7
Poorly/Well Organized	Floor Plan	17
	Furniture Arrangement	7
	Building Layout	2
	Management of BOQs	4
	Relative to Post	2
	Storage	4
	Built-ins/Personal Effects	4
	Other	12

diversity in the intrepretation of the bipolar descriptors, an ideal BOQ was rated virtually identically (there were no statistical differences) at the four Army installations. This fact adds support to the hypothesis that the 100-mm technique itself influences a person's response to a greater extent that the type of object being rated does.

4 DISCUSSION

This chapter discusses how the 100-mm technique can be used in facility evaluation and outlines areas requiring refinement before the technique can be accepted for more general use.

Possible Uses of Technique

As mentioned in Chapter 1, the *purpose for evaluating* a facility must be known in order to select the most appropriate methodology. If the intended *purpose for evaluation* is merely problem identification, the 100-mm technique is very adequate. The administration of a series of bipolar adjectives would then serve a very quick diagnostic purpose. At that level of generality, knowing what the specific problem may be with lighting is not important; the important thing is identifying *that there is a problem. Areas where no problems are* immediately identified would not need to be pursued further. Once a problem area such as lighting is identified, the level of necessary specificity would determine the next step. Either interviews or a series of specific test batteries could pinpoint that the problem might be control of lighting, type of lighting (natural or artificial), amount (too much or too little), location of switches, number of switches, glare, reflection, and so on.

The use of ideal ratings would, however, be meaningless as a diagnostic technique. With no anchor point (such as a set of ratings for "existing" features) for comparison, the ideal ratings are too ambiguous and have no well-defined reference. The results of this study indicate that there is very little discrimination of the concept of an "ideal" unless there is something to compare it with.

In terms of *what is being evaluated*, the ideal has been demonstrated to be an effective control variable both here and in previous CERL research. In a post-completion inspection (ER 415-3-11, par. 4), the existing conditions can be compared to the stable ideal through the administration of the bipolar scales before and after the 6-month occupancy requirement. The differences in the existing ratings for the two administrations could be compared to the stable ideal to measure the change.

On the basis of how *judgments are formed*, the ideal is strictly subjective. The internal standards people use to judge something as ideal vary in ways that are nearly impossible to quantify. The data do demonstrate that these ideal concepts normalize across subject pools to the point that, at least with the bipolar descriptors used here, their ratings vary only slightly, regardless of what is being evaluated.

Regarding *subject matter*, ideal scales paired with existing ratings can be used as a measure of change of judged quality. This usage would again require a before-and- after administration. These paired administrations can be used to rate facilities, building features, or occupant impressions, but the ideal rating cannot stand by itself.

Areas Requiring Refinement

A major part of the problem with the ideal scales used in this study is that the descriptors are ambiguous. Pairs of bipolar adjectives whose meaning and relevance are agreed upon by at least a majority of the subjects in the population to which they are to be applied are needed. The next step in refinement might be a rating of the relevance of the word pairs to be used. Word pairs that have an accepted meaning but are seen as irrelevant to the object being rated can only increase the variance and complicate any attempt to analyze the results. Once a set of commonly accepted relevant word pairs is found, attention could be refocused on the 100-mm technique rather than these superfluous contaminants. A pilot study could then be run to retest the discriminative ability of the 100-mm technique.

Perhaps the greatest problem with this study was that the 10 pairs of adjectives used were so general as to apply to almost any object. Words such as good-bad are almost universal in their applicability, especially given cultural response biases. The more general the word pair, the more likely it is to be applicable to many different objects, resulting in a lack of discriminability between them. This hypothesis is supported by the fact that the scales that did not discriminate between an ideal BOQ and an ideal wristwatch were those that were applicable to *both* BOQs and wristwatches: noisy/quiet, cluttered/uncluttered, run down/well kept, and poorly organized/well organized. On the other

hand, a scale such as brightly lighted/dimly lighted is much more applicable to a BOQ room than to a wristwatch; this difference is reflected by the fact that this scale discriminated quite well between the average ratings of the two.

5 CONCLUSION

The results of this study indicate that the ideal bipolar 100-mm rating scales could be a viable evaluation tool with one qualification: they must be paired with some dependent measure (such as existing scales) for them to provide any meaningful evaluation. Without such a basis of comparison, the scales do not discriminate between facility types well enough to provide meaningful input to facility evaluations. In fact, results indicated that the 100-mm technqiue may itself influence a person's response more than the type of object being rated does.

APPENDIX: DATA ANALYSES

Introduction

This appendix details the results of the analysis of the questionnaire data. Several analyses were undertaken, each having a different purpose in the understanding of different scaling issues. This appendix describes each of the analyses separately in the following sections.

The first section following this introduction describes the SPECTR program, discusses the results of the SPECTR analysis, and interprets the SPECTR analysis with implications for instrument revisions. The next section focuses on the scaling characteristics of the ideal items in distinguishing among different types of facilities. The final section presents results of an analysis comparing different locations within a given type of facility on the existing and ideal items.

SPECTR

A certain amount of invalid data is expected with any data collection effort. This has been especially true of data collected by questionnaire. Invalid data may come from several sources, such as subjects not actually attending to the questionnaire and therefore responding in a random fashion; subjects answering a few questions in a section and then quitting; subjects marking essentially the same response alternative to all items; subjects losing their places on the questionnaire, etc. In most cases, errors of this type are included in the analyses. However, with a new procedure–SPECTR –the more blatant forms of erroneous data can be eliminated, within certain statistical probabilities. The procedure was developed to screen out erroneous data from a 100-item questionnaire that surveys management practices and organizational climate.[5] In its present form, the procedure requires that the data meet the following specifications:

1. A high level of internal consistency, which, when combined with a questionnaire of sufficient length, would permit the separation of random responses from internally consistent responses

2. The positive end of the alternatives for each item assigned approximately randomly to the A and E end[6] of the alternative scale

3. The data scaled to five or fewer alternatives per item

4. The subscales of the instrument have near-zero intercorrelations. (*This consideration was not relevant* to the CERL data; i.e., no such scales exist, and deliberate positive and negative distortions which can be detected by this screen have limited applicability to CERL data.)

The CERL data were amenable to two of the screens available in this program: checking within-score consistency to eliminate random responses, and elimination of missing and illegal responses.

The screening procedures involved the following parameters:

1. Within-score consistency (WSC). Given a set of one or more score areas, each containing relatively homogeneous items, a score can be computed for each

---

[5] R. L. Ellison, C. Abe, D. G. Fox, and K. E. Coray, *Validation of the Management Audit Survey Against Employment Service Criteria* (U.S. Department of Labor, Employment and Training Administration, June 1976).

[6] J. C. Nunnally, *Psychometric Theory* (McGraw-Hill, 1967).

participant based on the average variance among the items in each score area. The average within-score variance is called the within-score consistency (WSC) measure for the set of score areas. A large WSC score indicates that the respondent did not respond in a similar fashion to items that had similar content and may have been marking responses on the questionnaire in a random fashion. An extremely low score indicates that the respondent was answering each item the same (e.g., all positive or all negative, or, if all the items have the same response scale, the respondent was simply marking all 10s or all 20s etc.). In short, there was very limited variability, suggesting that the respondent was deliberately distorting the data by, for example, consistently choosing the most negative alternative. The formula for the WSC measure per subject is:

$$WSC = \frac{\Sigma_{Scores}\left(\frac{\Sigma_{Items} X^2}{N_{Items}}\right) - \left(\frac{\Sigma_{Items} X}{N_{Items}}\right)^2}{N_{Scores}}$$

where X = item alternative value (reversed where necessary).

Σ across items is only for those items within a score. After the above score is computed for each subject, a frequency distribution is prepared for the subjects under study and a set of responses generated from a random number table. A cutting score is then set to eliminate the random response subjects, and the number of real subjects rejected is determined.

2. Missing and illegal responses. Although the CERL questionnaire was designed to accommodate questions with a range of 0 to 100, some responses fell outside the 0 to 100 range. This phenomenon was due either to faulty data preparation or to respondents' writing in a response greater than 100. Such a response is considered an illegal response. A large number of illegal responses would indicate that the respondent was not attending to the questionnaire, was not marking his/her responses in the appropriate area, or was deliberately making erroneous or random responses, etc. Further, a respondent may have a number of missing responses, i.e., items not responded to or left blank in the questionnaire. Missing responses indicate an unwillingness to cooperate, inadequate time to answer all the questions, absence for part of the administration of the questionnaire, etc.; the resultant scale scores would not adequately reflect the respondent's position on the dimension measured.

The dining hall questionnaire had the largest sample of participants (N = 534) and was selected for the SPECTR analysis. In these data, there were three relatively homogeneous subgroups of items available on which to base the SPECTR screens. These three subsets of items allowed the WSC score to be computed. The screening process used in computing the parameters and checking them against the cutoff levels involved asking the following questions:

1. Is the number of blank responses greater than 10 percent of the items being examined?

2. Is the WSC less than .20?

3. Is the WSC greater than 1.45?

If the answer to any of these questions was yes, all of the responses for the respondent were deleted from the file and were not included in further SPECTR analyses.

To set the cutting screens for the SPECTR run on the dining hall data, 200 cases of random data were created. Since the SPECTR program was created to work with items which ranged from 1 to 5, the CERL data (which ranged from 0 to 100) had to be rescaled. Unfortunately, rescaling is a relatively complex issue, since any rescaling procedure distorts the data in some way. Some procedures normalize the original data, other procedures tend to flatten the data, and still others tend to skew the data in a fashion not representative of the original distribution. For example, some logarithmic or exponential types of rescaling procedures may tend to make the original data appear to be curvilinear. All normalizing procedures lose the shape of the original data; the distortion may or may not be serious, depending upon the amount of skew. In the present case, a method of approximating the shape of the original distribution while still using the 1 to 5 data range was needed. The procedure used to accomplish this is discussed below.

To rescale the dining hall items, the overall across-facility means were calculated for each item; each case was then compared to that across-facility mean to obtain a standard score. The percentile of this standard score in a normal curve table was then determined. If the percentile was in the quintile from 0 to 20, the response was coded 1; from 21 to 40, the response was

coded 2; and so on. This procedure resulted in a very accurate approximation of the original raw data distribution. A frequency distribution of the rescaled data and a random sample of the original data was computed; the quintiles of the original data were closely duplicated in the rescaled data used in the SPECTR analysis, indicating that confidence should be placed in the rescaling procedure used.

One hundred random cases were created with a range of 0 to 100 and then rescaled in the manner described above. An additional 100 random cases were created with a range of 1 to 5. The random cases were created in two different ways to assess the quality of their randomness prior to setting the screens for the dining hall data. The cutting score on the WSC measure was set such that 5 percent of the random cases escaped the SPECTR screens. The results of applying this cutting score are shown in Table A1. This cutting score on WSC resulted in screening out 164 of the 534 cases for random response patterns. Another 15 cases were screened out by SPECTR for excess missing and illegal responses. One case was screened out for response set, i.e., very consistent responses. Thus, 180 (or 34 percent) of the 534 cases were screened out by SPECTR as invalid data. The most likely explanation for these results is that part of the data were based on random responses and the WSC procedure on the questionnaire was not sensitive enough to separate those respondents who answered somewhat inconsistently from the random answer cases.

Table A1
Percent of Total and Random Sample Data Passing SPECTR Screens

	Percent Passing Screens	
SPECTR Results	**All Cases**	**Random Cases**
Passing Screens	66	5
Failing Screens	34	95

To obtain additional information about this issue, means and standard deviations of the existing and ideal scales were computed on a sample of the data passing the SPECTR screens and the sample of cases that did not pass the screen (Table A2). Inspection of Table A2 indicates that the mean scores of the sample failing the SPECTR screens tended to parallel closely those of the passing sample. When a low score was obtained on the passing sample, a low score was also obtained by the failing sample and similar results were obtained for high scores. A priori, the failing sample, being based largely on random responses, would not be expected to have a pattern of consistency which approximated the sample that passed the SPECTR screens. Thus, the results obtained indicate that an important percentage of the sample failing the SPECTR screens were real data and not random cases, in spite of the fact that they resembled the random sample on the WSC measure.

Review of the standard deviations, however, indicates that there were marked differences in the two samples. The sample failing the SPECTR screens consistently had larger standard deviations and the differences were generally marked. This finding indicates that the SPECTR screens were working and that many of the subjects within the sample failing the SPECTR screens responded in a highly varied fashion to similar questions, approximating what would be expected with random responses.

The implications of these findings are that the SPECTR procedure apparently can be generalized to widely different kinds of data other than organizational climate measures which were constructed according to rigorous psychometric standards. However, for the screens to be effective on the dining hall data, additional internally consistent items and scores need to be generated. With the development of such internally consistent scores, the present results indicate that the sensitivity of the WSC measure would be increased and the random cases could be more accurately separated from real response cases. To obtain more information about the internal consistency of the dining hall data, additional analyses were carried out as described below.

Intercorrelations were computed for the objective satisfaction items, the semantic differential items which assessed the existing dining hall, and the intercorrelations among the semantic differential items for the ideal dining hall.

The data on the objective satisfaction items had an average item intercorrelation of .41 and an alpha coefficient[7] of .94. Since all of the items in this subset run the same direction, a condition not conducive to effective working of the SPECTR screens, all correlations were positive. This in effect made consistent answers easy to give, even though the subject may not have been reading the answers. A subject could merely answer a few questions, find the positive and negative end of the set of questions, and then proceed to answer

[7] J. C. Nunnally, *Psychometric Theory* (McGraw-Hill, 1967).

Table A2
Comparison of Means and Standard Deviations on Existing and Ideal Dining Hall Items for Samples Passed and Failed by SPECTR Screens

Variables	Means		Standard Deviations	
	Passing Sample	Failing Sample	Passing Sample	Failing Sample
Existing:				
Brightly/Dimly Lighted	50.73	47.45	24.47	35.00
Noisy/Quiet	65.98	72.38	22.83	30.18
Crowded/Uncrowded	30.79	27.05	26.72	32.96
Ugly/Beautiful	35.07	29.60	24.36	31.54
Drab/Colorful	59.67	61.61	28.85	36.41
Unpleasant/Pleasant	37.66	32.14	26.81	33.17
Uncluttered/Cluttered	38.12	37.79	25.50	34.85
Uninviting/Inviting	64.68	67.33	27.02	35.93
Run Down/Well Kept	52.79	47.80	27.70	35.42
Poorly/Well Organized	39.97	38.21	26.86	31.54
Ideal:				
Brightly/Dimly Lighted	46.95	50.84	26.76	36.21
Noisy/Quiet	24.97	24.91	22.00	32.39
Crowded/Uncrowded	74.78	73.03	23.84	33.00
Ugly/Beautiful	78.27	75.82	21.49	31.19
Drab/Colorful	20.00	23.91	22.03	32.83
Unpleasant/Pleasant	83.53	80.80	20.64	30.07
Uncluttered/Cluttered	75.11	68.82	27.58	36.48
Uninviting/Inviting	16.15	17.55	23.57	29.54
Run Down/Well Kept	14.02	12.97	19.94	22.51
Poorly/Well Organized	84.84	80.55	23.38	31.31

randomly on either the positive or the negative end depending on his/her general inclination. This subset of items, although subject to such a random response set, generally indicated no really deficient items, as most of the correlations were in the teens or considerably above, as indicated by the average correlation and the internal consistency results.

Table A3 shows the intercorrelations among the existing dining hall data. In contrast to the objective satisfaction items, the existing dining hall items, where some reversals were present, were considerably lower; in addition, some of the items are obviously not internally consistent with the set as a whole. For example, item 1 (brightly lighted/dimly lighted) has a pattern of essentially zero correlations with all the other items of the set, as does item 9 (unusual/usual). These results indicate that either these are ineffective items for assessing dining halls (i.e., they do not agree with the other items), or that they are assessing different domains of information and should be supplemented with additional items with which they would correlate and which would then boost the reliability of the areas being measured. Individual items often tend to be unreliable, with items placed in the first and latter part of a test booklet having only moderate intercorrelations. Furthermore, a slight restatement of an item may often result in a lower or different pattern of intercorrelations among supposedly similar items. Thus, as opportunities permit, subsets of items should be developed to measure internally consistent constructs which can be summed and interpreted to (1) simplify the presentation by dealing with scores instead of items; (2) obtain considerably more reliable measures; and (3) clarify the interpretations of the results.

Table A4 gives the intercorrelations among the semantic differential items for the ideal dining hall. In contrast to the previous table, the average item intercorrelations are higher; however, items 1 and 9 are still obviously either ineffective items or items that should be supplemented with additional measurement to form separate subscales.

Looking across these items, the intercorrelations among the items of the CERL data are somewhat lower than but do approximate the average intercorrelations of items in the Management Audit Survey of organizational climate for which the SPECTR screens were developed. However, the number of items is approxi-

Table A3
Intercorrelation Matrix of Variables from the Dining Hall Description Scale–Existing

Variables	1	2	3	4	5	6	7	8	9	10	11
1. Brightly/Dimly Lighted	–										
2. Quiet/Noisy	12	–									
3. Crowded/Uncrowded	–05	–29	–								
4. Ugly/Beautiful	–06	–38	31	–							
5. Colorful/Drab	08	31	–17	–55	–						
6. Unpleasant/Pleasant	–04	–38	32	68	–52	–					
7. Cluttered/Uncluttered	–04	–28	31	32	–21	39	–				
8. Inviting/Uninviting	03	40	–18	–59	58	–53	–24	–			
9. Unusual/Usual	–05	04	01	03	07	03	12	03	–		
10. Well Kept/Run Down	11	37	–14	–50	45	53	–28	51	00	–	
11. Poorly/Well Organized	–12	–31	26	45	–36	48	27	–35	16	–48	–

Table A4
Intercorrelation Matrix of Variables from the Dining Hall Description Scale–Ideal

Variables	1	2	3	4	5	6	7	8	9	10	11
1. Brightly/Dimly Lighted	–										
2. Quiet/Noisy	–02	–									
3. Crowded/Uncrowded	06	–66	–								
4. Ugly/Beautiful	03	–57	64	–							
5. Colorful/Drab	03	54	–55	–71	–						
6. Unpleasant/Pleasant	02	–62	62	77	–72	–					
7. Cluttered/Uncluttered	09	–36	41	41	–33	47	–				
8. Inviting/Uninviting	–03	58	–52	–66	69	–74	–39	–			
9. Unusual/Usual	–16	17	–15	–08	12	–10	–06	15	–		
10. Well Kept/Run Down	03	57	–57	–70	62	–74	–46	68	09	–	
11. Poorly/Well Organized	04	–48	52	58	–54	66	35	–55	00	–64	–

mately half, and thus, the somewhat lesser sensitivity of the SPECTR WSC measure to eliminate random cases is largely due to the lower number of items in the CERL data which went into the SPECTR program.

Considering all of the evidence available, the results suggest that an important percentage of the CERL data is probably completed randomly, but the actual percentage at this time cannot be determined. The results also indicate that these random cases could be effectively eliminated with the development of additional internally consistent items and subscores. In view of all the findings on SPECTR, no further work was carried out on the sample which failed the SPECTR screens. For the balance of the analyses in this report, the total sample of CERL data was used.

Comparisons of Different Types of Facilities on the Ideal Scales

An important question in considering the effectiveness of the existing and ideal sets of semantic differential items in evaluating different kinds of facilities is the extent to which the ideal items are sensitive to different kinds of facilities. That is, do the semantic differential items defining an ideal dining hall, an ideal BOQ, or an ideal wristwatch differ significantly? If the differences are trivial, then the set of items loses some credibility, as there are obvious differences in these objects. Figure A1 presents the ideal item means for the dining hall, BOQ, and wristwatch samples in graphic form. For each semantic differential item, two of the three different kinds of facilities are typically highly similar in their item means. On only a few items are

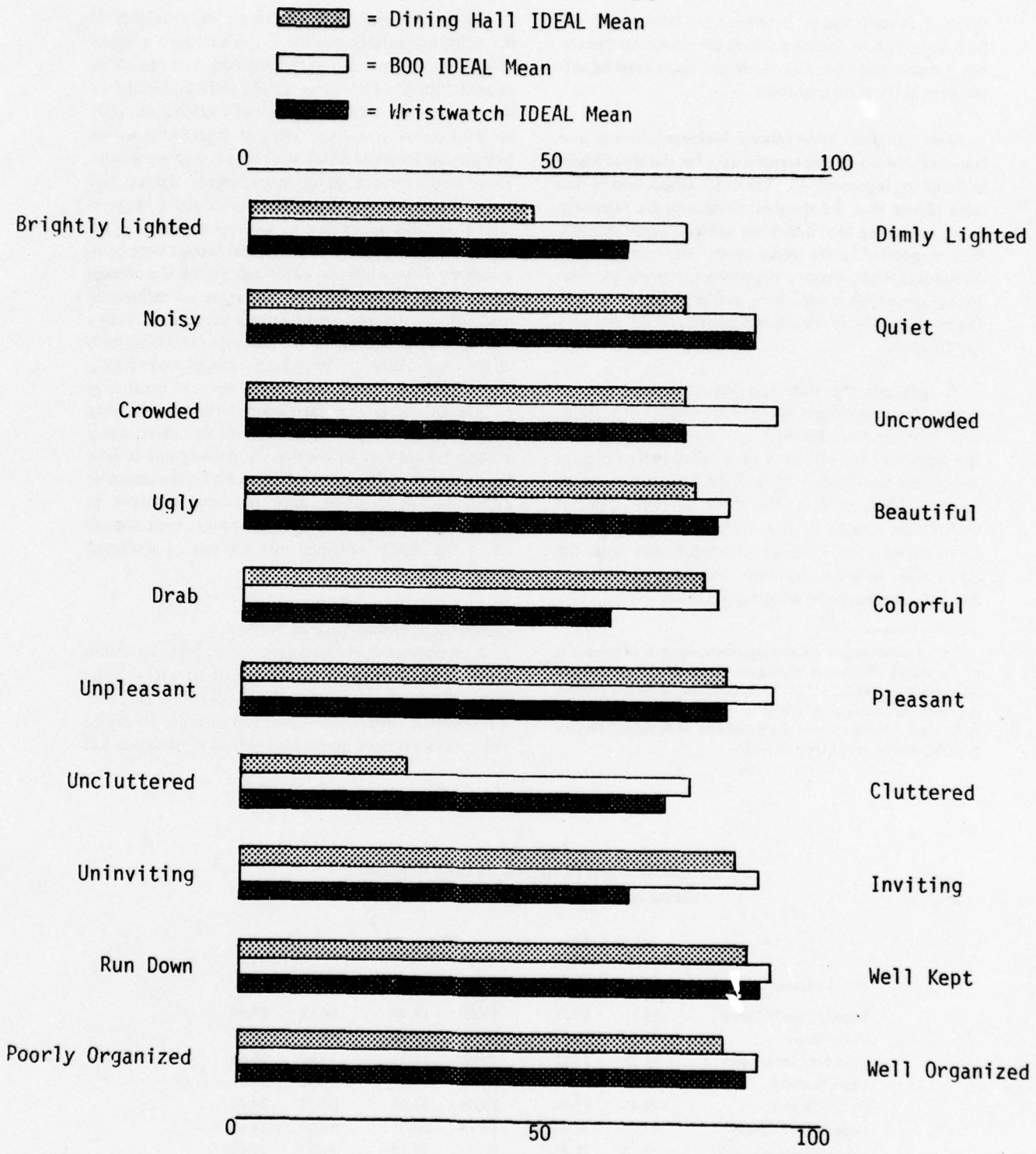


Figure A1. Comparison of the total sample ideal item means for the dining hall, BOQ, and wristwatch data.

there clear differences between the three kinds of facilities. Most of the item means are above .50 toward the positive end and only rarely are there marked differences in the item responses.

More detailed information, including means and standard deviations and sample sizes for the three kinds of facilities, is presented in Table A5. Inspection of this table shows that the standard deviations are relatively large, indicating that individual subjects varied in how they responded to the items on the 100-mm semantic differential scale. From a psychometric point of view, higher agreement would be a desirable result, in that firmer guidelines for the design of ideal facilities would be provided.

To examine the statistical significance of the differences between item means, intraclass correlations were computed on the three sets of means (Table A6). The intraclass correlation is an ideal statistic for comparing the significance of differences between means for these kinds of data. The three different types of facilities are treated as three different classes, and variations within the facilities in conjunction with the differences between facilities are used to examine the within versus between group variance.[8]

[8]More information concerning this statistic is presented in B. J. Winer, *Statistical Principles in Experimental Design* (McGraw-Hill, 1962), p 124; E. A. Haggard, *Intraclass Correlation and the Analysis of Variance* (Dryden Press, 1958); and R. L. Ebel, "Estimation of the Reliability of Ratings," *Psychometrika*, Vol 16 (1951), pp 407-424.

The key columns in Table A6 are the reliability of the individual ratings and the F's, which are a measure of significance of the results obtained. The results indicated that all of the items significantly differentiated between the three different kinds of facilities, but only on two items was there marked agreement within facilities and marked differences between group means. These items were brightly lighted/dimly lighted and cluttered/uncluttered. These two items could be questioned on other grounds, e.g., internal consistency. All of the other reliabilities of individual ratings were comparatively low. Although the reliability of the average ratings looks substantial, these averages are influenced markedly by the number of ratings per facility. With a relatively large sample of respondents describing each facility, reliability of the average ratings looks highly impressive; yet these must be interpreted cautiously because of the low results obtained for the reliability of individual ratings. Stated alternately, there was a substantial amount of overlap by participants in how they rated the different types of facilities; the semantic differential scale items were not very sensitive in producing marked differences in means, even though all of the results obtained met the test of statistical significance.

Comparisons Within Type of Facility

A number of different kinds of comparisons within types of facilities, i.e., dining halls, BOQ, etc., can be made. Previous research has already demonstrated that the semantic differential scales can measure perceived differences between an existing and an ideal dining hall

Table A5
Means and Standard Deviations on Ideal Items for the Dining Hall, BOQ, and Wristwatch Samples

	Dining Hall (N = 534)		BOQ (N = 287)		Wristwatch (N = 49)	
Variables	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.
Brightly/Dimly Lighted	48.25	30.26	73.55	19.40	64.08	23.94
Noisy/Quiet	75.34	25.52	86.97	12.38	86.71	18.71
Crowded/Uncrowded	74.20	27.22	79.96	18.71	73.98	26.66
Ugly/Beautiful	77.46	24.73	83.67	14.23	79.35	22.71
Drab/Colorful	79.46	25.08	80.08	14.64	63.55	24.02
Unpleasant/Pleasant	82.62	24.20	89.30	10.74	82.86	18.64
Uncluttered/Cluttered	29.76	31.21	76.57	25.14	72.53	27.01
Uninviting/Inviting	83.86	25.09	88.25	11.95	67.53	23.88
Run Down/Well Kept	86.66	20.16	90.52	10.41	88.55	18.30
Poorly/Well Organized	83.41	26.32	89.55	11.52	86.45	19.19

facility. Furthermore, the scales can also measure perceived differences in dining halls before and after extensive remodeling. Thus, there is no question that the semantic differential scales have some important advantages in assessing facilities. However, as pointed out earlier, they also have some problems and could be supplemented with additional measures to make a more effective measurement system.

In this section, a somewhat different perspective is used to make within-facility comparisons, e.g., the sensitivity of the semantic differential items in distinguishing between existing BOQs. Data on the existing and ideal items for each location are presented in graphic form. Comparisons of the existing and ideal item means on the BOQ data for each location are presented in Figures A2 through A5. The existing and ideal item means for each location are presented in Table A7.

These figures document the previous research findings that the semantic differential scales do produce marked differences between the existing and the ideal on each of the locations studied. In each case, the rating of the existing facility is considerably lower than that of the ideal facility on each of the semantic differential items.

An important question concerning these data is the extent to which the semantic differential scales can also differentiate among sites; that is, are the ratings of the existing characteristics of the sites significantly different across the different locations? To answer this question, intraclass correlations were computed on each of the semantic differential existing items to see if the item means of the various locations differed. The intraclass correlation results are presented in Table A8 for the existing and ideal items. The majority of the existing items were significant at either the .05 or the .01 level, although the reliability of the individual

Table A6
Intraclass Correlations on Ideal Items

Variable	Reliability of Average Ratings	Reliability of Individual Ratings	F df = 2, 835
Brightly/Dimly Lighted	.99	.27	78.272
Noisy/Quiet	.96	.11	27.719
Crowded/Uncrowded	.80	.02	4.917
Ugly/Beautiful	.86	.03	7.120
Drab/Colorful	.92	.05	12.175
Unpleasant/Pleasant	.89	.04	9.506
Uncluttered/Cluttered	.99	.54	239.653
Uninviting/Inviting	.95	.08	19.065
Run Down/Well Kept	.77	.02	4.278
Poorly/Well Organized	.85	.03	6.715

Table A7
Existing and Ideal Item Means for Each BOQ Location

Variable	Lewis		Bliss		Meade		Leonard Wood	
	Existing	Ideal	Existing	Ideal	Existing	Ideal	Existing	Ideal
Brightly/Dimly Lighted	58.36	64.82	51.46	74.55	42.47	74.11	44.33	75.69
Noisy/Quiet	58.09	85.09	49.37	87.84	48.30	88.51	56.26	84.05
Crowded/Uncrowded	50.88	79.94	58.61	81.55	44.38	79.47	47.16	77.90
Ugly/Beautiful	36.82	82.35	39.05	83.63	33.68	85.12	46.27	82.98
Drab/Colorful	35.56	76.71	35.84	80.29	26.88	81.36	41.23	80.33
Unpleasant/Pleasant	46.09	88.26	50.14	88.80	41.34	90.04	54.66	89.87
Uncluttered/Cluttered	49.24	76.74	55.63	71.59	44.97	81.05	52.24	78.74
Uninviting/Inviting	38.79	85.15	45.37	87.45	34.36	90.08	51.13	89.13
Run Down/Well Kept	49.71	90.91	52.32	89.99	45.43	91.19	60.08	90.36
Poorly/Well Organized	48.59	91.12	54.89	89.48	46.48	88.34	55.92	90.34

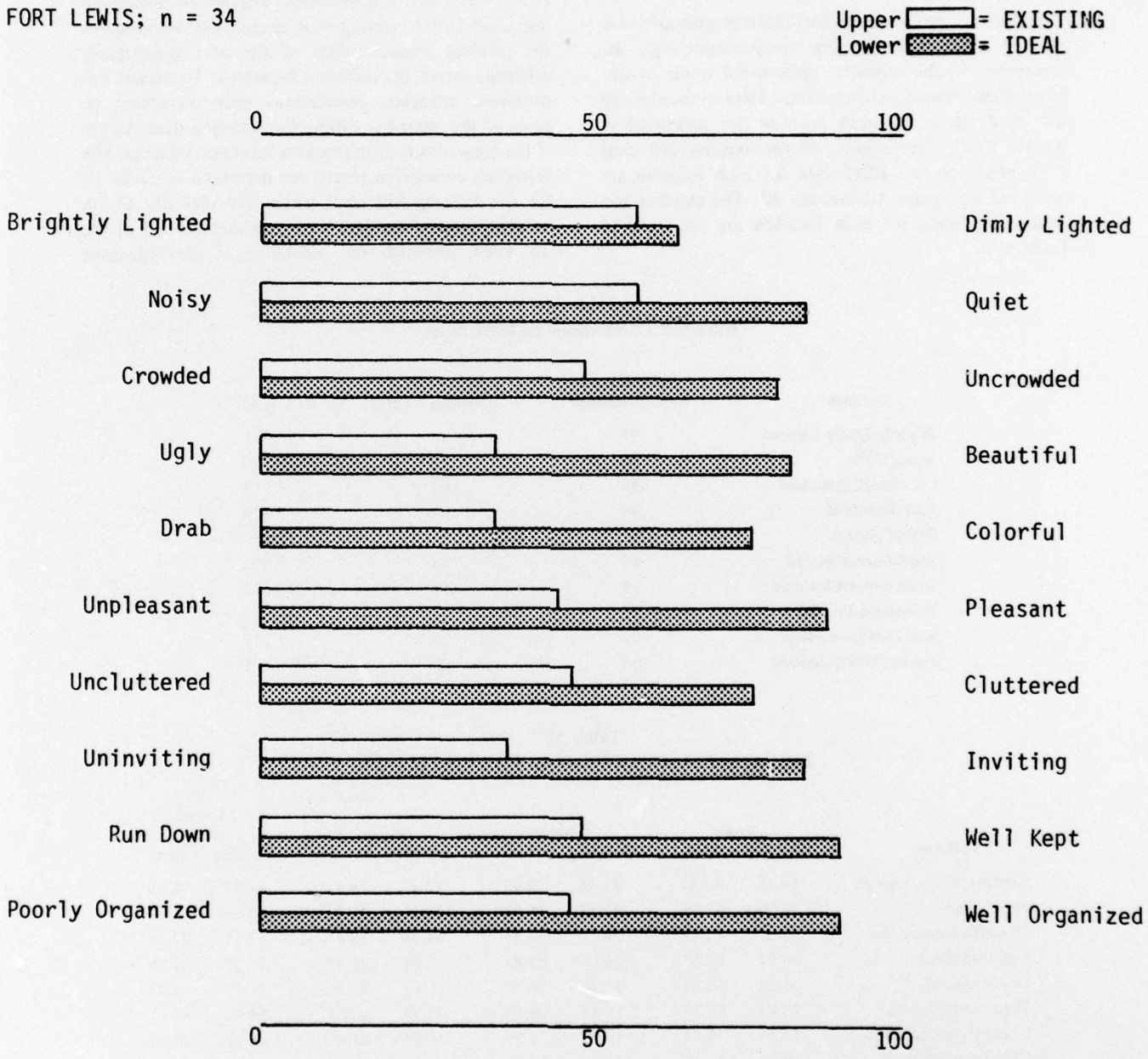


Figure A2. Comparison of the existing and ideal item means on the BOQ data–Fort Lewis.

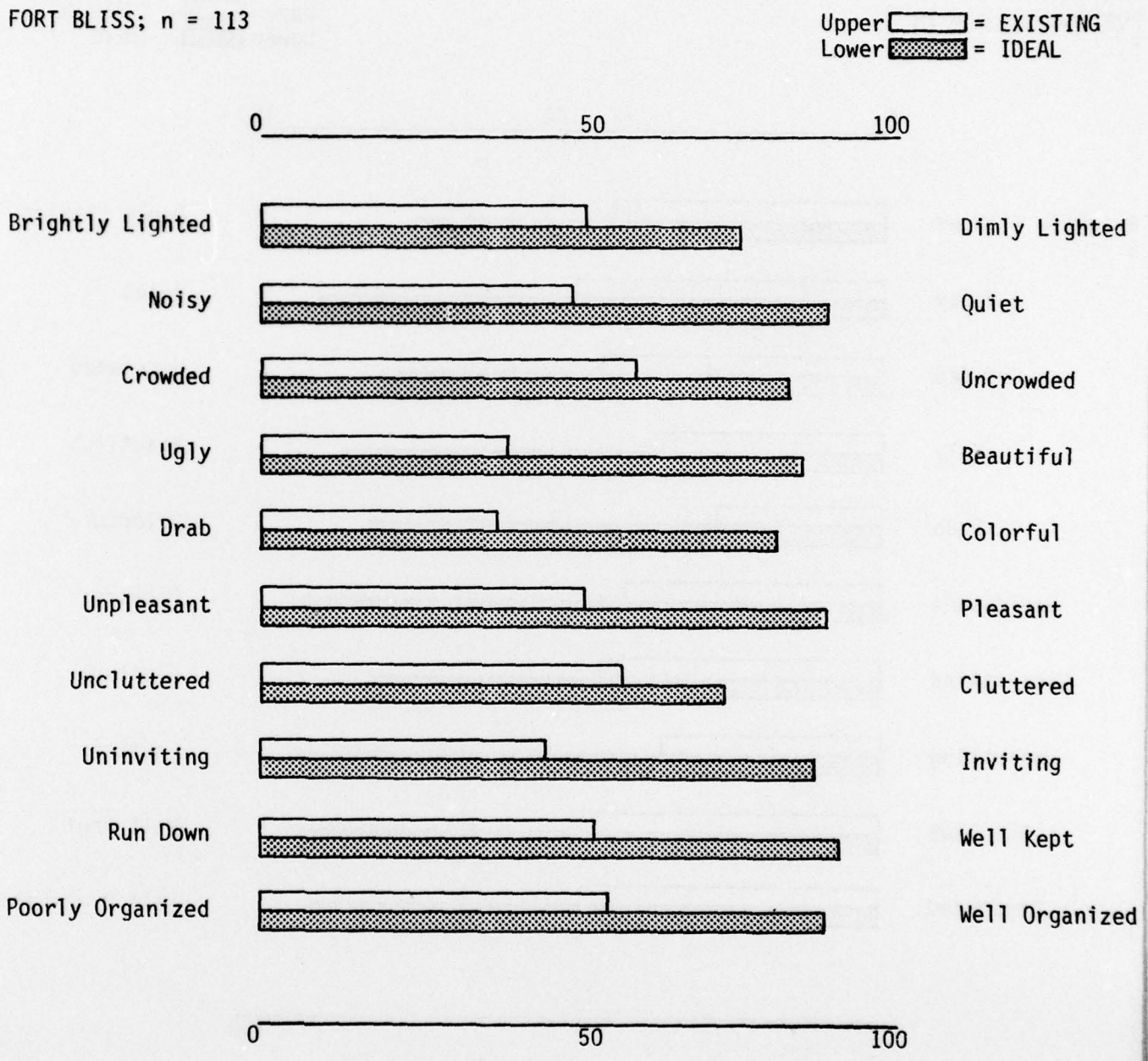


Figure A3. Comparison of the existing and ideal item means on the BOQ data—Fort Bliss.

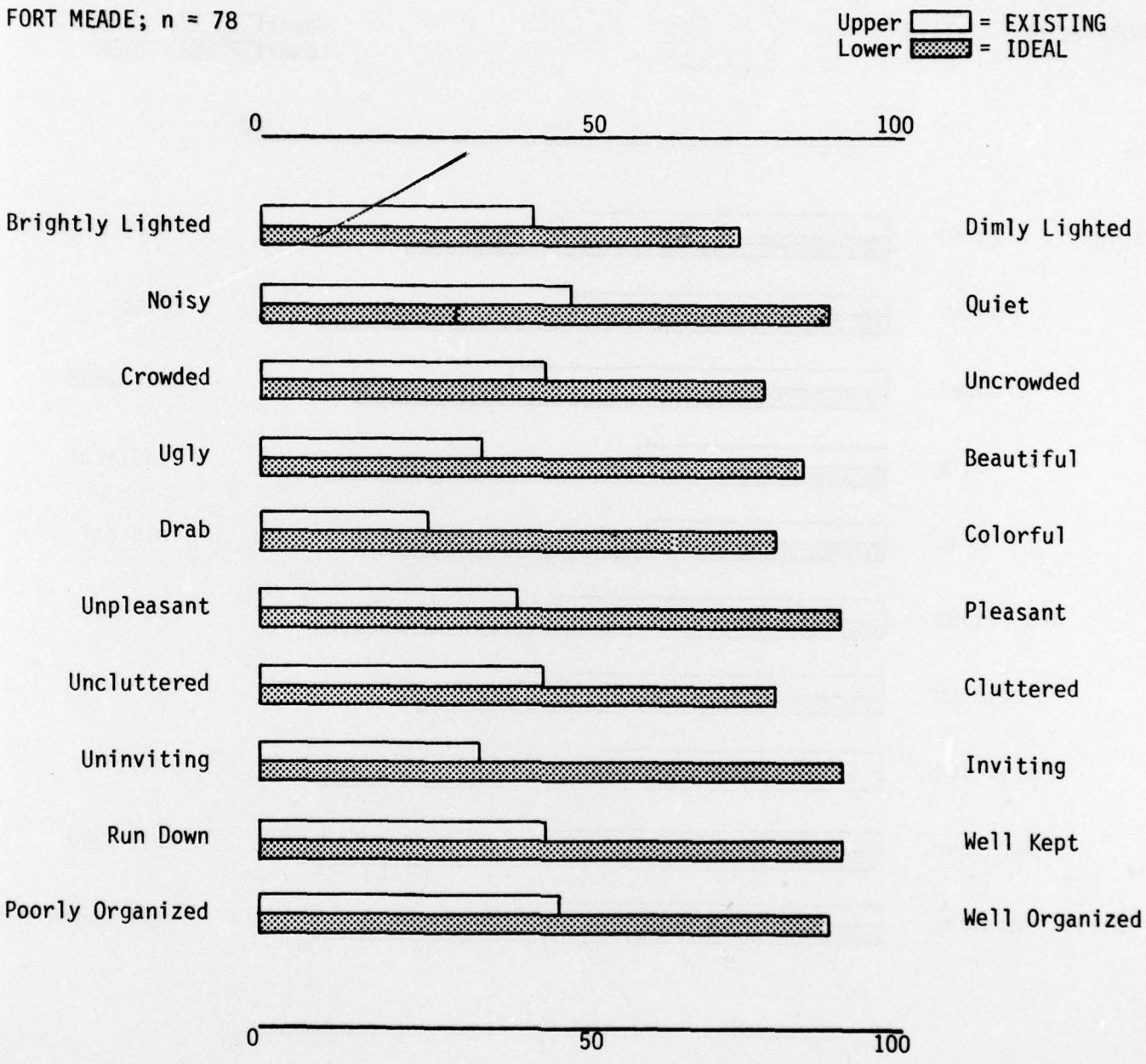


Figure A4. Comparison of the existing and ideal item means on the BOQ data—Fort Meade.

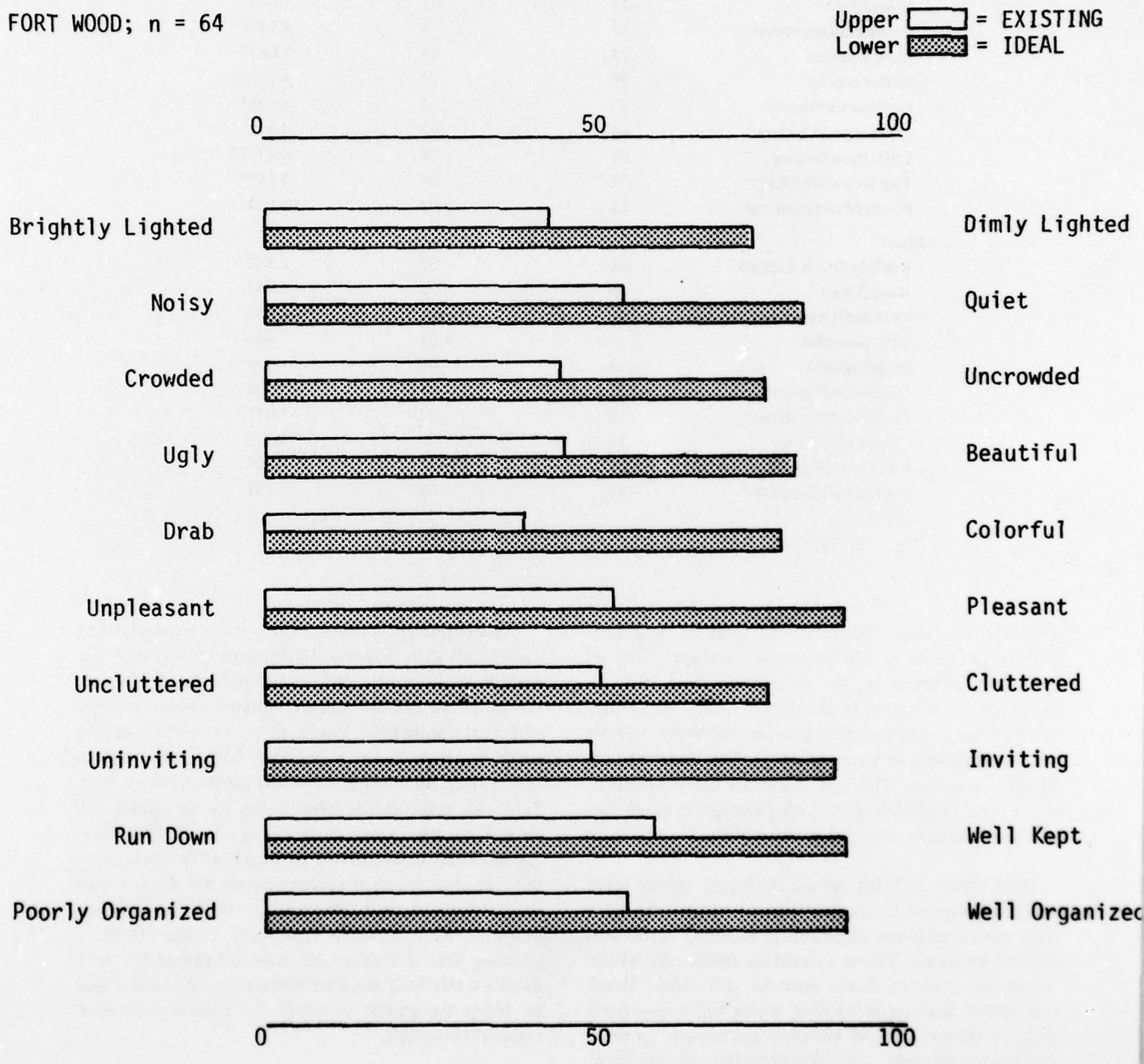


Figure A5. Comparison of the existing and ideal item means on the BOQ data—Fort Leonard Wood.

Table A8
Intraclass Correlations of BOQ Data

Variables	Reliability of Average Ratings	Reliability of Individual Ratings	F df = 3, 270
Existing:			
Brightly/Dimly Lighted	.77	.05	4.346**
Noisy/Quiet	.45	.01	1.807
Crowded/Uncrowded	.79	.05	4.798**
Ugly/Beautiful	.73	.04	3.667*
Drab/Colorful	.79	.05	4.657**
Unpleasant/Pleasant	.73	.04	3.698*
Uncluttered/Cluttered	.62	.02	2.615
Uninviting/Inviting	.85	.08	6.677**
Run Down/Well Kept	.72	.04	3.548*
Poorly/Well Organized	.62	.02	2.612
Ideal:			
Brightly/Dimly Lighted	.64	.03	2.780*
Noisy/Quiet	.36	.01	1.562
Crowded/Uncrowded	−.81	−.01	.551
Ugly/Beautiful	−1.27	−.01	.440
Drab/Colorful	−.25	.00	.798
Unpleasant/Pleasant	−2.27	−.01	.306
Uncluttered/Cluttered	.52	.02	2.071
Uninviting/Inviting	.36	.01	1.551
Run Down/Well Kept	−4.06	−.01	.198
Poorly/Well Organized	−.82	−.01	.551

**$p < .01$; *$p < .05$

ratings is low. These results indicate that the semantic differential items do differ across locations. This is important information, for it indicates that if total scores per location were developed across all of the items within a questionnaire, a normative base could be built indicating how a particular location compared to all other locations. This would provide useful information to the facilities engineer concerning the extent to which renovations, if any, were warranted.

Also shown in Table A8 are the results for the ideal items. In contrast to the existing item results, the ideal item means were not significantly different across the various locations. This is a desirable result, one which would be expected if the semantic differential items had utility. Ratings of the ideal dining hall do not need to be collected after an adequate norm base has been developed, because the characteristics of the ideal dining hall across locations are not significantly different.

Similar procedures were followed for the analysis of dining hall data. Figures A6 through A10 graphically present the semantic differential item means showing the items' capability to differentiate between existing and ideal dining halls. Again, in almost every case, for every location, there were clear differences between the existing and ideal items across the locations studied. Table A9 presents the item means for the dining hall data. These means were tested for significant differences via intraclass correlations (Table A10). The data again indicate that the existing item means did differ across the locations studied. The item means for the different locations were typically significant at the .01 level, showing that if total scores were developed for each facility a relatively sensitive measure could be developed to define the extent to which the particular location needed renovation.

For the ideal items, the general finding was again one of no significant differences between ideal dining halls

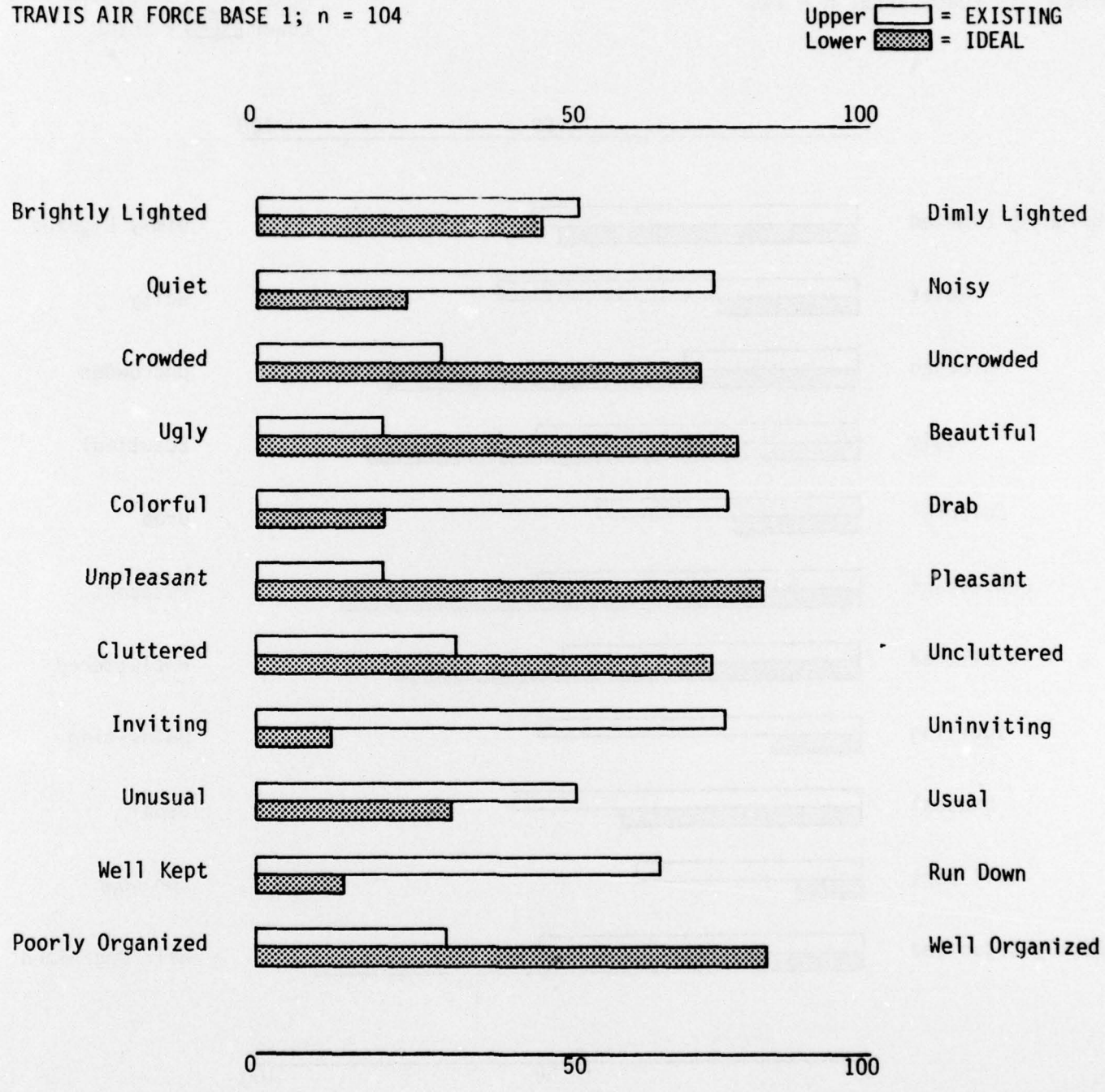


Figure A6. Comparison of the existing and ideal item means on Travis AFB Dining Hall 1 data.

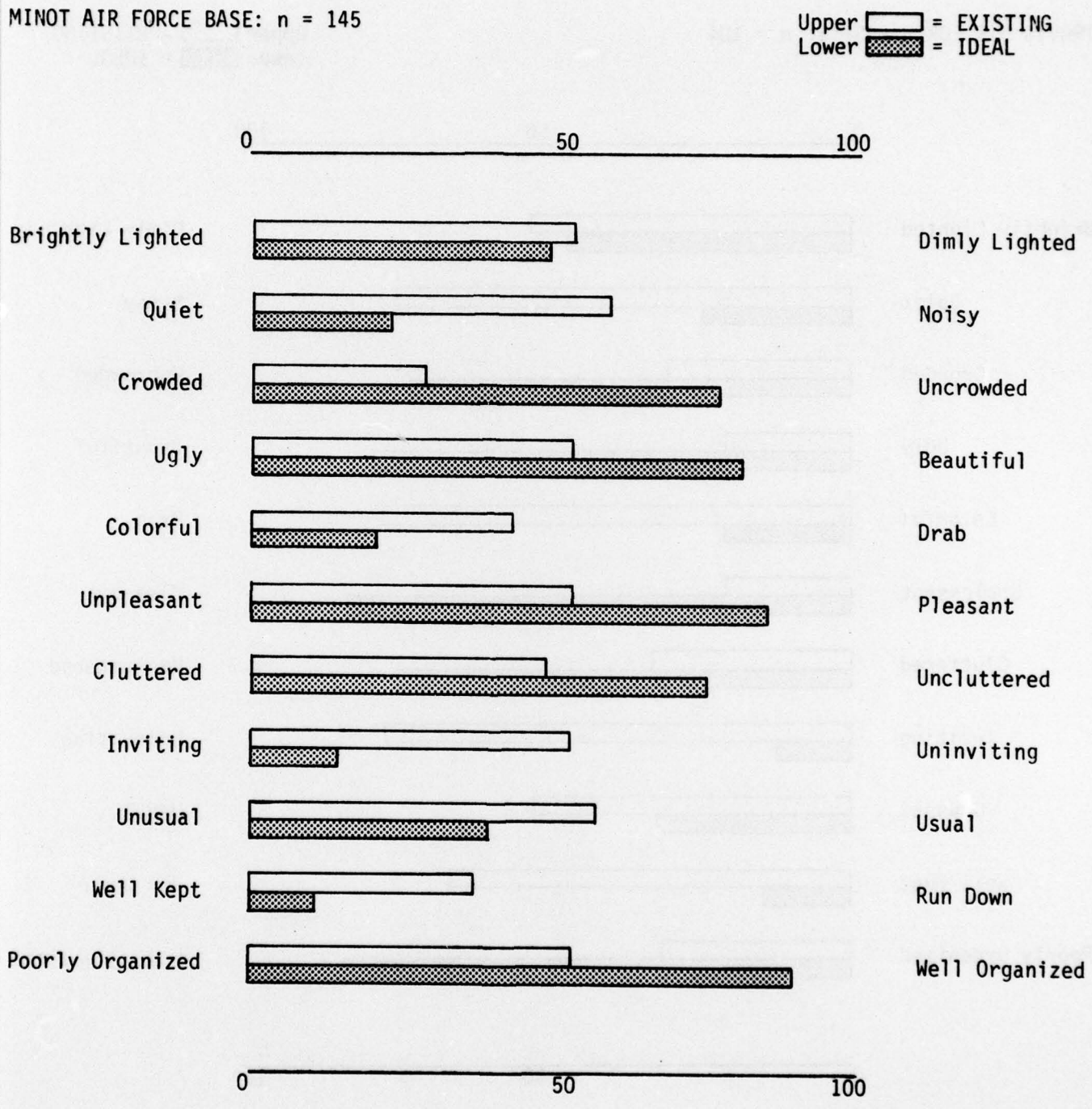


Figure A7. Comparison of the existing and ideal item means on Minot AFB dining hall data.

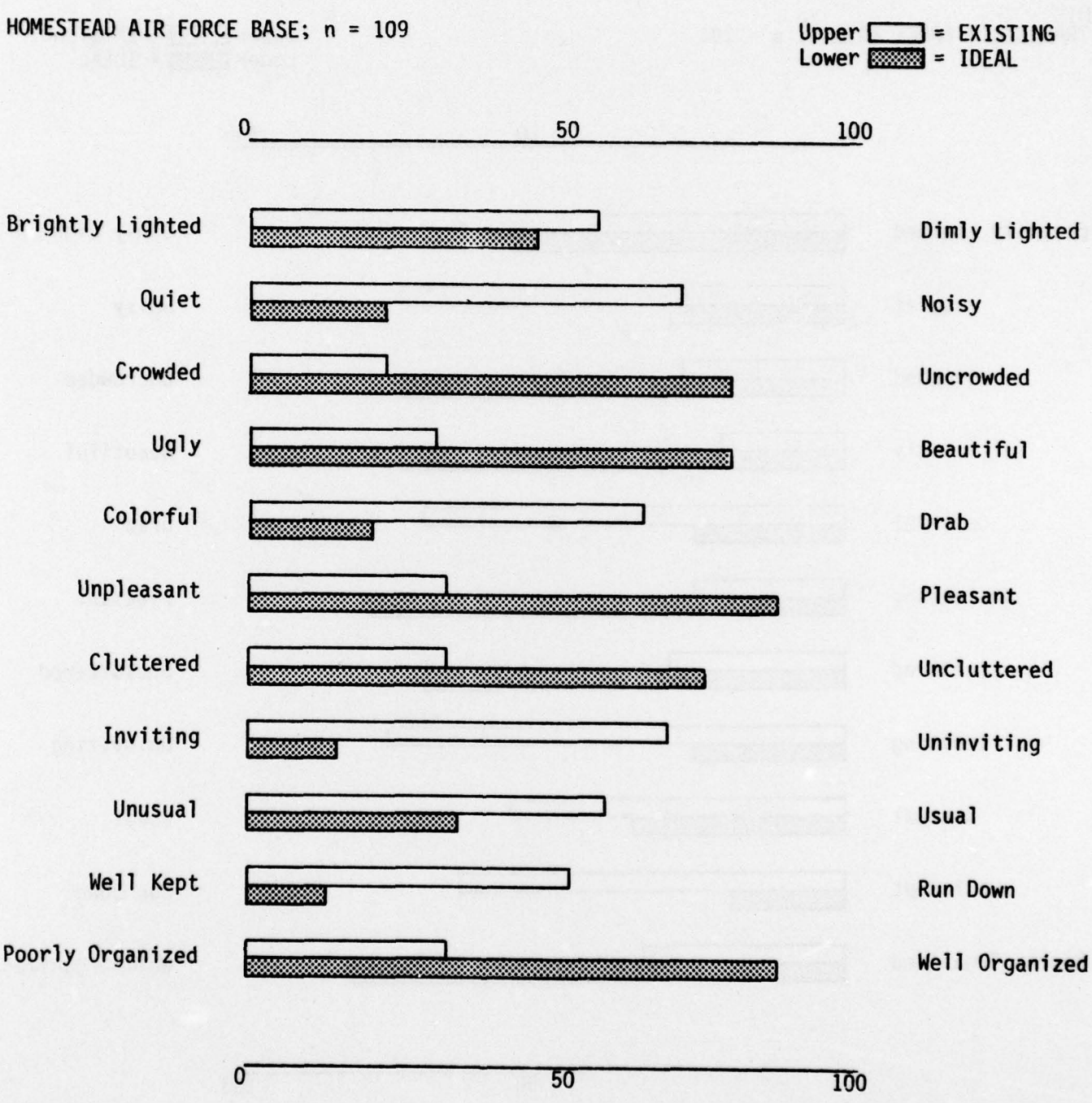


Figure A8. Comparison of the existing and ideal item means on Homestead AFB dining hall data.

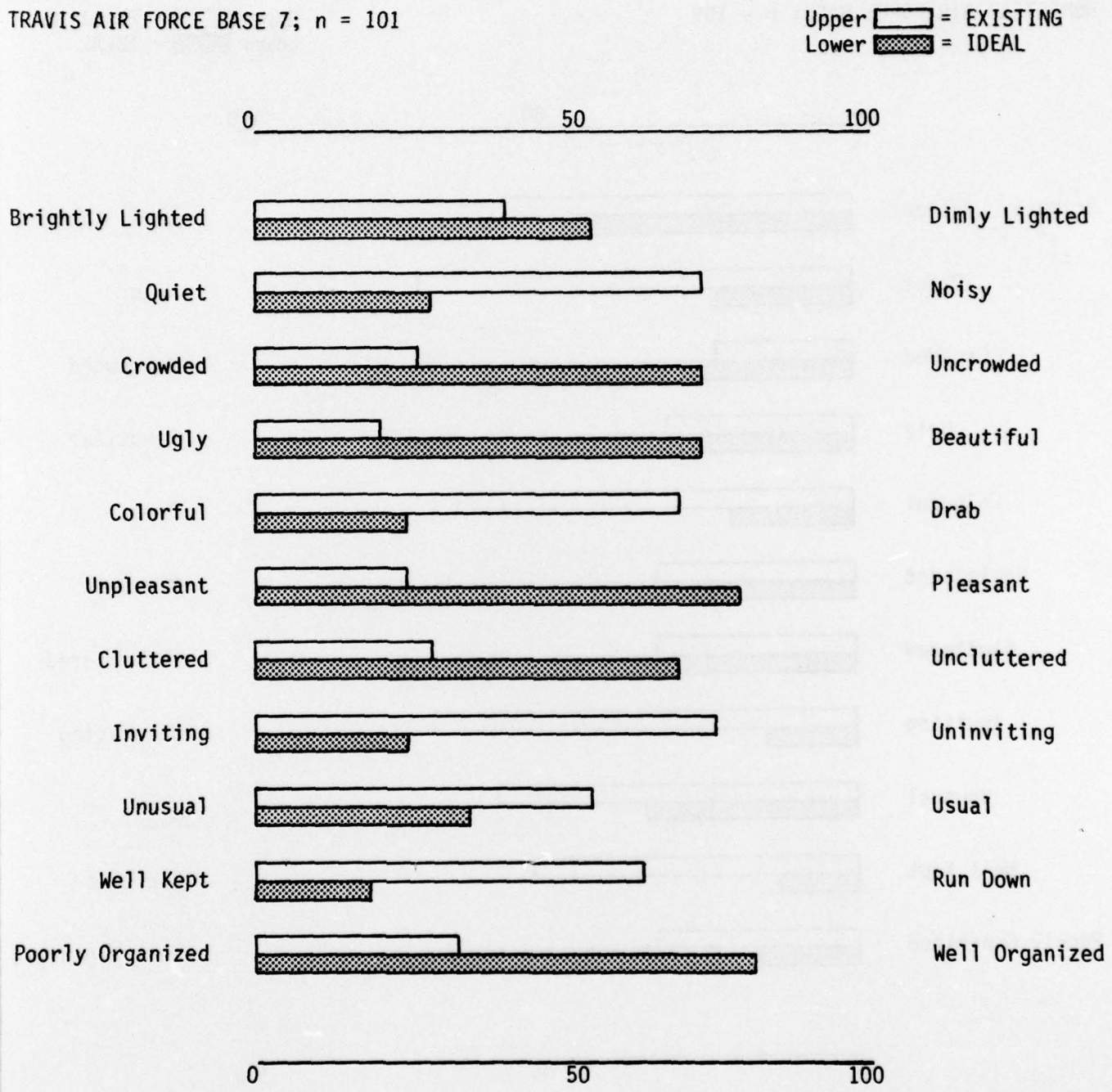


Figure A9. Comparison of the existing and ideal item means on Travis AFB Dining Hall 7 data.

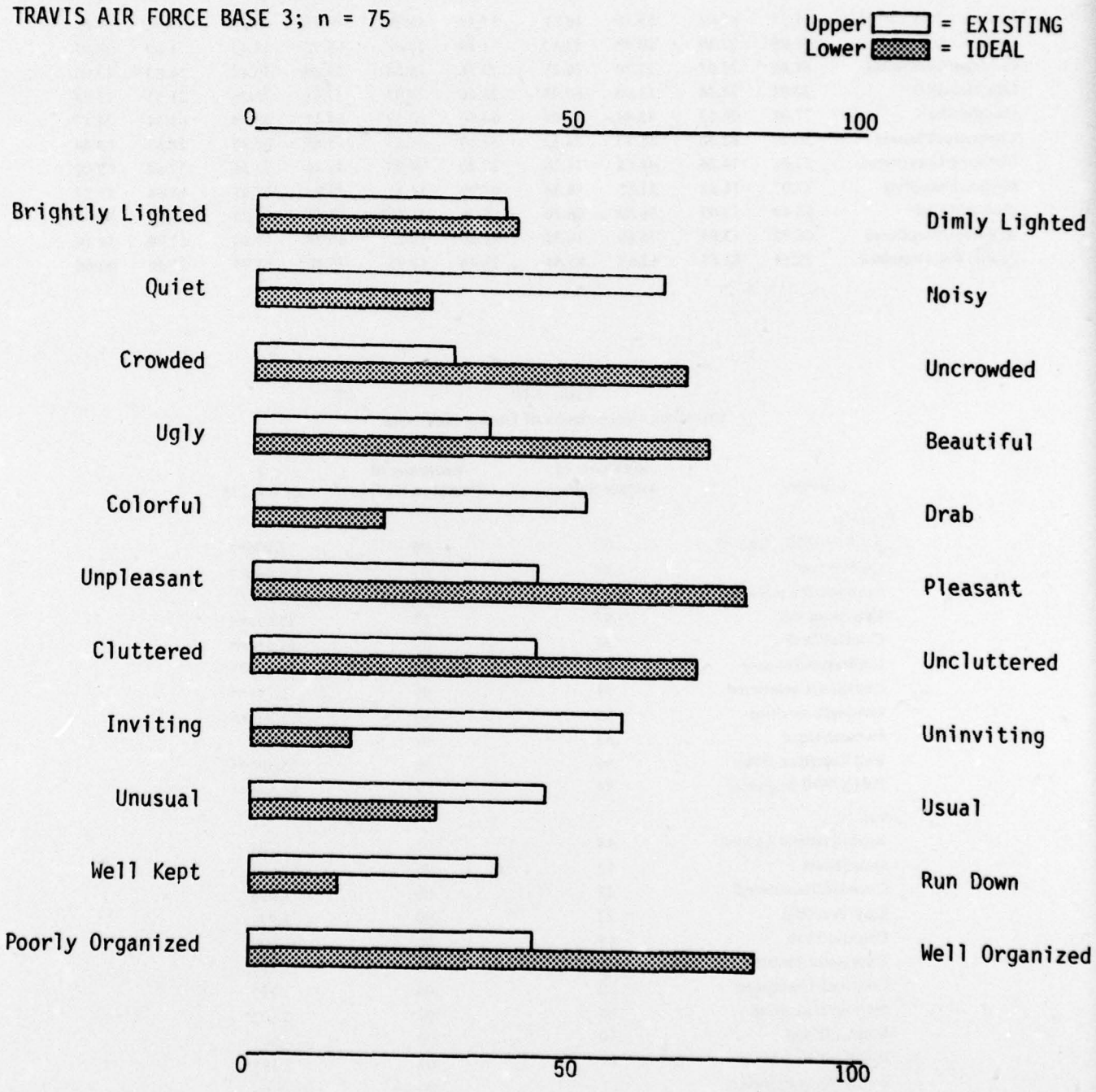


Figure A10. Comparison of the existing and ideal item means on Travis AFB Dining Hall 3 data.

Table A9
Existing and Ideal Means for Each Location–Dining Hall Data

Variable	Travis 1		Minot		Homestead		Travis 3		Travis 7	
	Existing	Ideal	Existing	Ideal	Existing	Ideal	Existing	Ideal	Existing	Ideal
Brightly/Dimly Lighted	51.95	47.02	53.50	48.81	57.42	46.29	40.86	43.27	39.82	54.67
Quiet/Noisy	75.09	23.88	59.02	23.17	71.19	22.88	65.92	28.43	71.60	28.37
Crowded/Uncrowded	31.46	72.07	33.70	76.25	22.36	78.00	33.00	70.47	26.83	72.05
Ugly/Beautiful	20.01	78.34	52.50	80.03	28.16	78.33	37.61	75.36	21.21	73.34
Colorful/Drab	77.05	20.52	43.44	19.64	63.80	20.39	54.12	22.74	68.24	24.43
Unpleasant/Pleasant	20.44	82.38	52.17	84.32	32.19	85.53	45.88	80.82	24.69	78.44
Cluttered/Uncluttered	33.06	74.24	48.08	74.20	33.30	74.99	45.84	72.15	27.64	68.42
Inviting/Uninviting	77.07	11.81	51.85	16.34	67.99	14.41	59.73	17.47	74.84	23.87
Unusual/Usual	53.49	33.09	56.62	38.36	58.76	35.46	47.65	30.03	54.31	34.07
Well Kept/Run Down	66.92	13.84	35.69	10.81	53.26	11.95	40.34	14.34	62.99	19.10
Poorly/Well Organized	29.51	81.65	52.17	85.44	33.44	85.93	46.20	81.95	32.48	80.66

Table A10
Intraclass Correlations of Dining Hall Data

Variables	Reliability of Average Ratings	Reliability of Individual Ratings	F df = 4, 525
Existing:			
Brightly/Dimly Lighted	.87	.06	7.906**
Quiet/Noisy	.88	.07	8.346**
Crowded/Uncrowded	.67	.02	3.065*
Ugly/Beautiful	.97	.27	39.674**
Colorful/Drab	.96	.18	23.644**
Unpleasant/Pleasant	.97	.22	30.150**
Cluttered/Uncluttered	.91	.09	11.155**
Inviting/Uninviting	.94	.12	15.827**
Unusual/Usual	.35	.01	1.530
Well Kept/Run Down	.96	.20	26.700**
Poorly/Well Organized	.94	.12	15.599**
Ideal:			
Brightly/Dimly Lighted	.44	.01	1.784
Quiet/Noisy	.13	.00	1.151
Crowded/Uncrowded	.29	.00	1.406
Ugly/Beautiful	.22	.00	1.276
Colorful/Drab	−.65	.00	.608
Unpleasant/Pleasant	.29	.00	1.399
Cluttered/Uncluttered	−.32	.00	.757
Inviting/Uninviting	.68	.02	3.102*
Unusual/Usual	−.10	.00	.907
Well Kept/Run Down	.61	.02	2.588*
Poorly/Well Organized	−.11	.00	.905

**$p < .01$; *$p < .05$

across the locations studied. Stated alternately, once a comparatively large sample of ratings on the ideal dining hall was obtained, collecting more data would no longer be necessary, since the ideal means of dining halls across the locations studied did not differ.

To summarize, the intraclass correlation results of the existing and ideal semantic differential items support the further use of these items. The existing items successfully discriminated between different locations on both BOQs and dining halls. Equally important, there were no significant differences among the ideal items on either dining halls or BOQs. However, the level of sensitivity of the existing items in discriminating between different locations was relatively low. A total score across the existing items would increase the sensitivity of the items in measuring characteristics of facilities. Other kinds of items could be expected to be more sensitive in discriminating the unique characteristics of the facilities at each location and thus would be a valuable supplement to the semantic differential items.

REFERENCES

Brauer, R. L., and D. L. Dressel, *Concepts for the Generation, Communication, and Evaluation of Habitability Criteria*, Special Report D-78/ADA-041187 (U.S. Army Construction Engineering Research Laboratory [CERL], 1977).

Dressel, D. L., et al., *Army Family Housing: Preferences and Attitudes about Housing Interiors, Vol III: Predictors of Satisfaction with Housing Interiors*, CERL Technical Report D-48/ADA011187 (CERL, April 1975).

Gibbs, W., *Comparative Study of Consumer Attitudes at Three Air Force Dining Facilities*, Interim Report D-40/ADA000711 (CERL, 1974).

Gibbs, W., *Comparison of Consumer Satisfaction Before and After Dining Facility Renovations at Travis AFB, CA*, Technical Report D-28/AD784056 (CERL, 1974).

Ebel, R. L., "Estimation of the reliability of ratings," *Psychometrika*, Vol 16 (1951), pp 407-424.

Ellison, R. L., C. Abe, D. G. Fox, and K. E. Coray, *Validation of the Management Audit Survey Against Employment Service Criteria* (U.S. Department of Labor, Employment and Training Administration, June 1976).

Haggard, E. A., *Intraclass Correlation and the Analysis of Variance* (Dryden Press, 1958).

Nunnally, J. C., *Psychometric Theory*, (McGraw-Hill, 1967).

Winer, B. J., *Statistical Principles in Experimental Design* (McGraw-Hill, 1962).

CERL DISTRIBUTION

Chief of Engineers
ATTN: DAEN-ASI-L (2)

Dept of the Army
WASH DC 20314

Defense Documentation Center
ATTN: DDA (12)
Cameron Station
Alexandria, VA 22314

Veneklasen, Wayne D
Use of "ideal" ratings as a standard for evaluating facilities / by Wayne D. Veneklasen, Roger L. Brauer, Bruce Sevy. -- Champaign, IL : Construction Engineering Research Laboratory ; Springfield, VA : available from National Technical Information Service , 1978.
37 p. ; 27 cm. (Special report -- Construction Engineering Research Laboratory ; E-132).

1. Buildings-evaluation. 2. Semantic differential technique. I. Brauer, Roger L. II. Sevy, Bruce. III. Title. IV. Series: U.S. Construction Engineering Research Laboratory. Special report ; E-132.